Desenvolupament d'un software per a l'estudi de dades de l'Institut Català de la Salut

Universitat Politècnica de Catalunya (UPC)

Facultat d'Informàtica de Barcelona (FIB)

Treball final del Grau en Enginyeria Informàtica

Especialitat en computació

Autor: Martí Zamora Casals

Director: Ricard Gavaldà Mestre (Dept. LSI, UPC)

Codirectora: Sílvia Cordomí Saborit (Institut Català de la Salut)

Quadrimestre de primavera 2014

Agraïments

Vull agrair al professor Ricard Gavaldà per haver-me proposat aquest projecte i tot l'interès que hi ha posat durant aquests mesos. També vull agrair a la Sílvia Cordomí, les doctores Juliana Ribera, Ester Amado, Esther Limón i a José Luís del Val tota l'ajuda prestada i tot l'interès que han tingut en el projecte des del primer moment.

Índex

1 RESUM....4
1.1 Català....4
1.2 Castellano....4
1.3 English....5
2 INTRODUCCIÓ....6
2.1 Objectius....8
2.2 Impacte social, ambiental i econòmic....9
2.3 Estat de l'art....10
3 DEFINICIÓ DEL PROJECTE....13
4 PLANIFICACIÓ I PRESSUPOST....16
4.1 Pressupost....17
5 PRELIMINARS....19
5.1 Dades disponibles....19
5.2 Elements de Data Mining....21
6 ALGORISMES I TÈCNIQUES DE MIRERIA DE DADES....26
6.1 Generació de grafs de malalties i medicaments....26
6.2 Generació de regles d'associació....27
6.3 Detecció d'episodis no tancats i medicaments sense explicació....29
7 IMPLEMENTACIÓ....31
7.1 Servidor....31
7.2 Mòdul d'anàlisi....32
7.3 Interfície....35
8 Experiments, resultats i avaluació....43
8.1 Regles d'associació....43
8.2 Grafs....47
9 TREBALL FUTUR....50
10 CONCLUSIONS....54
11 BIBLIOGRAFIA....55

1 RESUM

1.1 Català

Aquest projecte és una col·laboració entre l'Institut Català de la Salut (ICS) i la Universitat Politècnica de Catalunya (UPC). Ha consistit en desenvolupar diverses tècniques per tal de poder explotar les dades de què disposa la base de dades de l'ICS. Una part del projecte ha consistit en elegir quines tècniques de mineria de dades es podien utilitzar i després s'han implementat per tal de posar-les a prova.

Les dades amb que s'ha treballat, corresponen als centres d'atenció primària de l'ICS a Barcelona, que tenen una població assignada de 1,6 milions de persones, i comprenen prop de 12 milions d'actes assistencials ("visites"). Altres estudis similars que s'han fet no comprenen una població tan variada i tan gran.

Concretament s'ha treballat en una manera de relacionar malalties i medicaments, és a dir, que un programa informàtic pugui aprendre quins medicaments es donen per a cada malaltia i detectar aquells pacients que tenen malalties que no es corresponen amb els medicaments que tenen receptats. El resultat és una eina que poden utilitzar experts en medicina per tal d'explotar les dades de forma automàtica des d'una interfície.

Com que es tractava d'una prova pilot, han quedat moltes qüestions i portes obertes que s'hauran d'anar responent amb investigació posterior tant en el camp mèdic com en el camp de les ciències de la computació.

1.2 Castellano

Este proyecto es una colaboración entre l'Institut Català de la Salut (ICS) i la Universitat Politècnica de Catalunya (UPC). Se ha basado en el desarrollo de varias técnicas para poder explotar los datos de que dispone la base de datos del ICS. Una parte del proyecto se ha basado en elegir que técnicas de minería de datos se podrían utilizar y después se han implementado para ponerlas a prueba.

Los datos con que se ha trabajado, corresponden a los centros de atención primaria del ICS en Barcelona, que tienen una población asignada de 1,6 millones de personas, y que comprenden cerca de 12 millones de actos asistenciales ("visitas"). Otros estudios similares que se han hecho no comprenden una población tan variada y tan grande.

Concretamente se ha trabajado en una forma de relacionar enfermedades y medicamentos, es decir, que un programa informático pueda aprender que medicamentos

se recetan para cada enfermedad y detectar los pacientes que tienen enfermedades que no se corresponden con los medicamentos que tienen recetados. El resultado es una herramienta que pueden usar los expertos en medicina para explotar los datos de forma automática desde una interfaz.

Al tratarse de una prueba piloto, ha quedado muchas cuestiones y puertas abiertas que se tendrán que responder con investigación posterior tanto en el campo médico como en el campo de las ciencias de la computación.

1.3 English

This project is a collaboration between the Catalan Health Institute (ICS) and the Polytechnic University of Catalonia (UPC). It has consisted in developing several techniques to mine the data from the ICS data base. A piece of the project has been based in electing which techniques of data mining could be used and then they have been implemented in order to test them.

The data, corresponds to the ICS first aid centers in Barcelona, which have an assigned population of 1.6 million people, and include near 12 million health care events. Other similar studies do not comprise such large and varied population.

Specifically, we have worked in a way to relate diseases and medicines, i.e. a software that can learn which medicines are prescribed for every illness and detect the patients that have diseases that do not match with the medicines that have been prescribed. The result is a tool that can be used for medicine experts to mine the data automatically from an interface.

As this was a pilot, so many questions remain open and they will need to be answered with further research in the medical field and in the computer science field.

2 INTRODUCCIÓ

El sistema sanitari és un dels pilars fonamentals de la societat del benestar. Al mateix temps és, juntament amb l'educació, una de les despeses més grans per a l'administració pública.

S'ha comprovat que les persones amb diverses malalties cròniques consumeixen la major part dels recursos sanitaris. S'observa que el 5% de la població consumeix el 50% dels recursos tant farmacèutics com hospitalaris i d'assistència primària. A més, l'envelliment de la població ha fet augmentar en els últims anys la despesa sanitària. En els propers anys es preveu que aquesta tendència s'accentuarà encara més. Tot això, fa necessari que aquest sistema s'hagi d'optimitzar al màxim tant des del punt de vista econòmic com també des del punt de vista de la salut i la qualitat de vida dels pacients. La Figura 1 i la Taula 1 ens mostren respectivament la freqüentació del sistema sanitari per edats i la despesa en farmàcia a la ciutat de Barcelona.

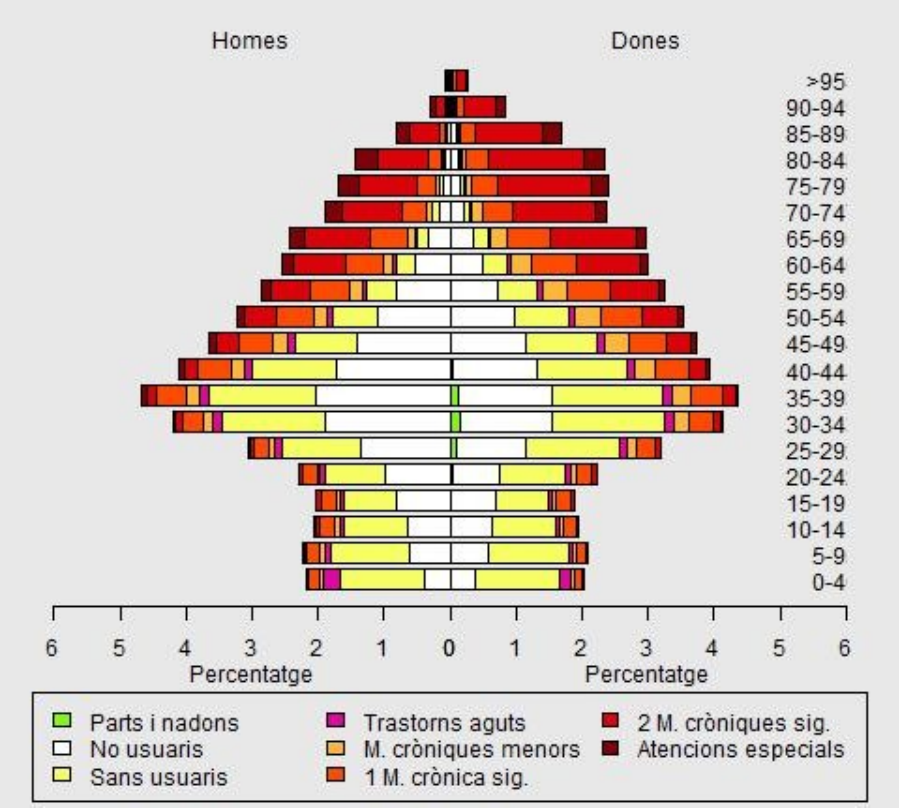


Figura 1. Freqüentació del sistema sanitari per franges d'edats a Barcelona, any 2013. [Font: Mòduls per al seguiment d'indicadors de Qualitat, Divisió d'Anàlisi de Demanda i d'Activitat. Servei Català de la Salut. Departament de Salut.]

Despesa de farmàcia. Euros per habitant segons Agrupació amb Clinical Risk Groups							
	Nivell de gravetat						
	1	2	3	4	5	6	Total
Sans usuaris	11	40					12,2
Trastorns aguts	18,8	94,7					46,5
1 M. crònica menors	66,5	140,5					84,3
M. cròniques menors	98,2		184,3	298,5			167,8
1 M. crònica sig.	143	236,2	484	435,2	402,8	1.560,10	185,1
2 M. cròniques sig.	348	551	723,2	847,5	1.025,00	1.291,70	558,7
M. cròniques dominants (3+)	913	1.009,60	1.221,90	1.331,20	1.466,60	1.716,90	1.212,30
Neoplàsies complexes	325,1	517,1	848,3	1.175,40	1.090,90		823,4
Necessitats sanitàries elevades	226	227,1	635,4	1.100,60	712,2	1.676,80	557,8

Taula 1. Distribució de la despesa sanitària per habitant. Els "Clinical Risk Groups" són un estàndard de classificació de la gravetat de pacients [http://www.gencat.cat/ics/memoria_2011mb/files/assets/basic-html/page6.html] [Font: Mòduls per al seguiment d'indicadors de Qualitat, Divisió d'Anàlisi de Demanda i d'Activitat. Servei Català de la Salut. Departament de Salut.]

Paral·lelament, la digitalització dels sistemes d'informació durant els darrers anys permet tenir a disposició dels investigadors grans quantitats de dades per poder-les analitzar. Les dades sanitàries no són una excepció i, malgrat les dificultats de coordinació, des de fa cinc anys es disposa de dades uniformitzades per als centres públics. Tot i això, la gran quantitat de dades fa necessària l'existència d'eines especialitzades per a l'extracció d'informació d'aquestes dades.

El projecte sorgeix arran de converses inicials entre el grup de recerca LARCA de la UPC i l'Institut Català de la Salut, d'ara endavant ICS, sobre la possibilitat d'aplicar tècniques de mineria de dades a l'anàlisi de dades del sistema sanitari català. El director del projecte, Ricard Gavaldà, és coordinador del grup LARCA i la co-directora, Sílvia Cordomí, és adjunta a Gerència d'Atenció Primària de l'ICS. Es proposa com una prova pilot, enfocat a dades de diagnòstic i de prescripció de medicaments en l'àmbit primari, no hospitalari. Té, per tant, un cert caràcter exploratori.

Participants del projecte

Apart del Ricard Gavaldà i la Sílvia Cordomí que són el director i la codirectora respectivament en aquest projecte també hi han treballat altres persones associades a l'ICS: la Dra. Juliana Ribera com a coordinadora i gestora del projecte entre la UPC i l'ICS, les doctores Ester Amado i Esther Limón com a expertes en farmàcia i medicina respectivament i el José Luís del Val que ens va proporcionar les dades i ens va explicar l'estructura de la base de dades.

2.1 Objectius

El projecte consisteix en desenvolupar un software capaç d'absorbir un subconjunt de les dades de la base de dades de l'ICS i, mitjançant tècniques de *data mining* i *machine learning*, obtenir-ne informació útil per als investigadors i responsables.

Es busca que el software produeixi resultats que permetin als investigadors descobrir noves informacions. Per fer-ho els algorismes s'han de poder aplicar sobre tot el conjunt de pacients a la vegada sense fer hipòtesis però que al mateix temps es puguin filtrar les dades de manera que un cop es tenen indicis d'alguna novetat es puguin filtrar les dades i quedar-se amb la part que interessa per poder fer un estudi concret.

Es disposa de les dades corresponents als centres d'atenció primària de l'ICS a Barcelona Ciutat durant l'any 2013. Principalment les que fan referència als problemes de salut i les prescripcions de cada pacient. Tot i això, el software hauria de poder escalar per utilitzar dades d'altres anys i altres zones.

Les funcionalitats principals que es van pensar per al software inicialment, principalment en converses amb el director i la codirectora, van ser:

1. Caracterització de grups (clústers) de problemes de salut. Això significa obtenir els grups de malalties que s'acostumen a trobar conjuntament a determinats pacients o trobar les malalties secundaries que pivoten al voltant d'una malaltia principal.

2. Caracterització de grups de receptes de medicaments i relació amb els grups del punt 1. Significa trobar quins medicaments són receptats en els grups de l'apartat anterior i sota quines condicions.

3. Capacitat de detectar “outliers” i fer anàlisis diferencials, per exemple, districtes, hospitals, metges que tenen un comportament molt diferent a la resta alhora de receptar o dels diagnòstics que s'hi produeixen, així com malalties o grups de malalties que es mediquen de manera no uniforme (amb dos, tres, o més grups de medicació) i que caldria estudiar per què. Es farà aprofitant els resultats dels punts 1 i 2.

4. Presentació de les dades per tal de poder-les visualitzar de forma entenedora i poder-hi realitzar experiments i consultes de forma interactiva. Les dades d'entrada han de poder ser filtrades per l'usuari de manera que pugui fer experiments més complexos i selectius que simplement abocar totes les dades al programa.

Com explicarem més endavant, converses amb altres investigadors de l'ICS van permetre refinar, enfocar, i reconduir lleugerament els objectius inicials. Es tracta que l'eina desenvolupada sigui útil i al mateix temps fàcil d'utilitzar per usuaris experts en medicina i farmàcia però no necessàriament amb coneixements avançats d'informàtica o bases de dades, tenint a més en compte la durada d'un treball de final de grau (4 mesos), i que es pretén continuar el projecte de col·laboració entre la UPC i l'ICS en el futur.

Un requisit clar des del principi, és que els mètodes emprats havien d'escalar. Malgrat que les dades de què hem disposat durant el projecte no han estat excessivament grans (en el sentit que no han presentat problemes de temps o memòria), si volem que el software pugui escalar amb dades d'altres anys i zones, hem hagut de tenir en compte que hi ha alguns mètodes de mineria de dades i aprenentatge automàtic que no escalen bé amb la mida de les dades.

2.2 Impacte social, ambiental i econòmic

No hi ha una manera clara de quantificar l'impacte que pot suposar el projecte i menys encara sense conèixer els resultats que es trobaran. Clarament, l'impacte no es donarà com a conseqüència directa del projecte a curt termini, sinó que vindrà donat per investigacions que puguin derivar-ne posteriorment i l'ajut al descobriment de coneixement pràctic a investigadors, clínics i planificadors del sistema sanitari. Malgrat tot, podem pensar en l'impacte social, econòmic i ambiental que podria tenir indirectament el projecte.

Impacte social

Per a la vessant social, es podrien trobar maneres de millorar la qualitat de vida dels pacients a conseqüència de la millora de la medicació que reben. Especialment la gent gran i els malalts crònics són els que poden veure's més beneficiats per una millora d'aquest tipus, ja que la seves malalties condicionen el seu dia a dia. Per mesurar la qualitat de vida dels pacients hi ha indicadors com els Clinical Risk Groups (CRG), tot i això, mesurar-ho seria responsabilitat dels metges del ICS.

També es podria reduir la hiperprescripció, és a dir, prescriure als pacients més medicaments dels que són necessaris. La hiperprescripció comporta que els pacients experimentin efectes secundaris que moltes vegades es tapen fent ús d'altres medicaments augmentant així, el risc de patir altres efectes secundaris.

Impacte econòmic

Es podria aconseguir una gestió més racional i efectiva dels recursos que aporta l'administració pública a la sanitat, cosa que permetria alleugerir l'augment del cost que es preveu en els següents anys a causa de l'envelliment de la població. Per exemple, només que s'aconseguís un estalvi d'un 10% en els medicaments que consumeixen el 1% dels pacients crònics s'aconseguiria un estalvi anual total de 1,5 milions d'euros.[1] Per tant, ajudem a fer el sistema sanitari públic més sostenible.

Impacte ambiental

Una millor gestió dels medicaments pot comportar un estalvi d'aquests i, per tant, que es generin menys residus de medicaments, que són complicats de reciclar.

2.3 Estat de l'art

S'han fet diverses propostes sobre aplicar mineria de dades a les dades sanitàries sobretot a Estats Units de cara a explotar els Electronic Health Records (EHR) que són uns registres que integren els registres sanitaris dels pacients [1][2].

Per suposat, no és el primer cop que es porten a terme investigacions sobre conjunts de pacients i se n'estudien tant les relacions de problemes de salut entre ells i amb altres factors (demogràfics, ambientals) i medicació. Podem dividir aquests estudis en dos grups: els estudis concrets i els estudis genèrics. En citem alguns dels molts que segurament podríem trobar, per destacar la novetat (creiem) d'aquest.

Estudis concrets

En els estudis concrets s'investiga sobre una malaltia concreta o grups de malalties. Malgrat ser estudis concrets, en tots els casos s'explica que els mètodes poden ser aplicats a d'altres malalties per exemple. Però s'han anomenat concrets perquè cada estudi s'enfoca a una sola malaltia de forma individual.

Alguns exemples podrien ser, intentar predir si els inhibidors Cox-2 poden causar infarts a pacients amb determinades característiques [3].

En aquest cas [4] s'aplica *enrichment analysis* per detectar característiques en els pacients que els permeti distingir els que han patit un determinat esdeveniment de la malaltia com podria ser aquells pacients que han patit una mort sobtada. Aquest estudi

1 Font: Indicadors IMP. IMP21: Despesa en receptes (en euros per habitant) Població global de Catalunya. Any 2012. ICS.

només es fa sobre pacients que pateixen miocardiopatia hipertròfica. D'aquesta manera volen intentar predir quan un pacient té risc de patir un determinat esdeveniment de la malaltia. En el mateix estudi també dades genètiques.

L'estudi [5] troba malalties relacionades amb la hèrnia paraesofàgica mitjançant regles d'associació a partir de les malalties que tenen els pacients que tenen la hèrnia paraesofàgica. A més, també ho combinen amb el coneixement expert dels metges. Aquest estudi s'assembla, tècnicament, al nostre ja que farem servir també variants de regles d'associació i intentarem, al menys, facilitar la combinació amb coneixement expert.

A diferència d'aquests treballs, el nostre projecte vol fer un estudi genèric, no centrat en una malaltia sinó obert a trobar relacions interessants i significatives entre qualsevol conjunt de malalties.

Estudis genèrics

En els estudis genèrics, es busca descobrir patrons sobre tot el conjunt de pacients sense fer hipòtesis concretes. Aquests estudis serveixen com a orientació per poder fer estudis posteriors concrets i més exhaustius. És a dir, serveixen per trobar relacions que poden haver passat inadvertides com podria passar si fixéssim una hipòtesi d'entrada, i fer-les visibles als experts per a un estudi més detallat.

Aquest és el cas de [6]. Que, mitjançant diversos mètodes de mineria de dades, intenten trobar noves associacions entre malalties a partir dels EHR de 667.000 pacients. És possiblement l'estudi més ambiciós quant a volum de pacients a nivell mundial.

També s'han fet estudis on es busquen maneres de crear xarxes de malalties i d'aquesta manera poder clusteritzar-les, és a dir, agrupar aquelles malalties que apareixen alhora en un pacient. N'és un exemple el treball [7], on s'usen els EHR de 327.000 pacients.

Altres estudis s'han basat en buscar la co-ocurrència de parells de malalalties en pacients i descobrir quines són associacions reals a partir d'estadístiques i utilitzant heurístiques per detectar les associacions que són realment interessants [8]. Al projecte *The Human Disease Network* [9] es va crear un graf d'aquelles malalties que estaven relacionades entre elles i al mateix temps, lligar-ho amb aquells gens que hi estan relacionades a partir d'una llista d'uns 1300 problemes de salut.

La novetat del plantejament del nostre treball (excloent de moments aspectes tècnics) és doncs:

• No s'ha fet mai cap estudi genèric (i pocs d'específics) de co-ocurrència de malalties i medicaments a Catalunya, segons l'ICS.

• No hem trobat cap estudi que abarqui aquesta quantitat de pacients potencials (gairebé 1,7 milions) ni de registres d'assistències (més de 10 milions).

D'altra banda, fins i tot concretar les preguntes a formular i les respostes a obtenir a partir programa ha estat ha estat un repte per si mateix, ja que fins ara dins de l'ICS mateix no havia considerat amb deteniment les possibilitats ni quin profit voldria treure d'aquestes dades.

3 DEFINICIÓ DEL PROJECTE

Aquest projecte és d'un caràcter molt exploratori, i per això es va considerar una metodologia no gaire convencional i diferent d'altres projectes de construcció de software on els requisits són tancats des de molt aviat. En concret, es va preveure fer el treball en quatre fases principals:

1. Determinació de requisits i funcionalitats
2. Preprocessament de les dades, anàlisis manuals i tria de mètodes a usar
3. Automatització: construcció del software
4. Experimentació i validació dels resultats

Totes han estat portades a terme amb la planificació prevista, excepte la última que no s'ha acabat de completar per els motius que s'hi expliquen.

Etapa 1: Determinació de requisits i funcionalitats

Consisteix en acabar de decidir conjuntament amb els metges i responsables de l'ICS quines funcionalitats havia de tenir el software tenint en compte les restriccions tecnològiques i les restriccions de temps que té un treball final de grau.

Les funcionalitats que s'han implementat finalment han sigut 4 principalment:

• Capacitat de generar grafs que relacionin les malalties entre elles per veure quins conjunts de malalties es donen freqüentment en un mateix pacient, fer el mateix amb els medicaments.

• Poder generar regles d'associació entre malalties i medicaments (i a l'inversa) per saber a quin percentatge de pacients amb una malaltia se'ls dona un medicament concret. A més les regles d'associació serveixen per les funcionalitats següents.

• Detectar episodis de malalties tancats. Aquesta funcionalitat va sorgir arran de les converses amb les expertes de l'ICS. Ens van demanar trobar una manera de poder detectar quan a la base de dades consta una malaltia per un pacient que realment ja no la té. Això passa per exemple quan un pacient va a un ambulatori amb una malaltia com pot ser una grip. El metge li fa unes receptes (que tenen data de caducitat). El pacient es troba bé i no torna a l'ambulatori per fer el seguiment. Les receptes caduquen i ja no consten com a actives, però la grip segueix constant com a malaltia activa. Aquesta nova part va ser inclosa en el projecte, perquè semblava important abans d'aplicar altres tipus

d'anàlisis. No havia sigut prevista inicialment i a causa d'això es van reduir en part els objectius de visualització que quedaran per millorar en projectes posteriors.

• Detectar pacients amb medicaments que no tenen explicació. Aquesta funcionalitat també va sorgir arran d'una reunió. Per cada medicament que pren un pacient intentar trobar una malaltia que expliqui aquest medicament, és semblant al punt anterior però al revés. Ens permet detectar quan en un pacient se li estan donant medicaments sense una explicació.

Podem veure que el primer apartat correspon als punts 1 i 2 dels objectius, el segon correspon al punt 2 i el tercer i el quart corresponen al punt 3 dels objectius, ja que en certa manera s'estan detectant pacients que no segueixen la norma.

Etapa 2: Preprocessament de les dades, anàlisis manuals i tria de mètodes a usar

Durant aquesta fase es van:

• Preparar les dades per eliminar-ne d'errònies, completar les mancants, recodificar a fi que fossin més fàcils de donar a eines de mineria de dades, etc.

• Es van efectuar una sèrie d'anàlisis manuals (bé: amb eines de mineria de dades estàndard) per comprovar les potencialitats de diversos mètodes d'anàlisi i fer-nos una primera idea de la seva escalabilitat

• Finalment es van elegir quins mètodes de mineria de dades s'utilitzarien i quins de nous caldria implementar.

Etapa 3: Automatització, construcció del software

En aquesta part s'ha implementat una versió estable que integra i automatitza els mètodes triats durant la etapa anterior. El software és capaç d'extreure les dades, netejar-les i organitzar-les, filtrar-les tal com l'usuari cregui convenient, i aplicar els mètodes d'anàlisi rellevants. Per exemple, s'han de poder descartar determinats pacients o malalties que compleixin una certa condició.

En aquesta part també s'ha afegit una interfície gràfica que permet a l'usuari tant configurar la tasca com visualitzar-ne els resultats de manera intuïtiva.

Avancem que la solució per la qual s'ha optat ha sigut desenvolupar un servidor de Node.js que permet connectar-se als usuaris de forma remota o en servidor local. Quan un usuari es connecta des del seu navegador es crea un procés de Java que és

l'encarregat d'executar totes les comandes que són transmeses des del navegador de l'usuari. D'aquesta manera es permet que el servidor pugui estar per exemple, a prop de la base de dades per poder-ne extreure la informació i realitzar els càlculs des d'allà. Addicionalment aquesta arquitectura permet que els usuaris puguin compartir experiments i resultats de manera senzilla.

Bàsicament el software permet executar els algoritmes que porten a terme les funcionalitats escollides en l'etapa 1 i visualitzar-ne els resultats en forma de graf o en forma de taula. També es poden descarregar els fitxer en format CSV o GML en el cas dels grafs per poder-ho visualitzar amb software extern com pot ser Excel o Gephi.

La interfície permet seleccionar subconjunts dels resultats d'un experiment com poden ser les regles relacionades amb una determinada malaltia o una part del graf centrant-lo en un node i a una certa profunditat.

Etapa 4. Experimentació i validació dels resultats

La última fase es consisteix en una bateria d'experiments per obtenir i, en la mesura del possible, comprovar els resultats. Aquests resultats també seran validats estadísticament per assegurar que per exemple no s'han extret conclusions amb massa poques dades o que no són artefactes estadístics dels mètodes usats.

Els resultats s'analitzaran juntament amb els experts de l'ICS, que diran quins resultats ja eren coneguts i quins són inesperats i per tant aporten nou coneixement, cosa que podria dur-los a prendre decisions en la seva operativa o bé a obrir noves línies d'investigació.

Aquesta part no s'ha pogut portar a terme per raons de calendari: el software requereix un cert temps d'aprenentatge, i l'anàlisi de resultats per part dels experts de l'ICS més temps és més laboriós potser del previst inicialment. Relativament aviat es va veure que aquesta part hauria de quedar fora del calendari d'aquest treball de final de grau.

4 PLANIFICACIÓ I PRESSUPOST

El projecte va començar a principis de gener i havia d'acabar a mitjan juny. Es va planificar el seu desenvolupament en les quatre fases explicades al capítol anterior, i a més s'hi afegeix el seguiment del mòdul de Gestió de Projectes (GEP) que transcorre paral·lelament al projecte i una fase final de redacció de documentació i presentació del treball.

Fase assignatura GEP

Consta dels lliuraments següents: definició de l'abast, planificació pressupost i sostenibilitat, presentació preliminar, recopilació del context i la bibliografia, plec de condicions de l'especialitat, document final i presentació.

Va durar des del 17 de febrer fins el 14 de març.

Determinació de requisits i funcionalitats

Aquesta primera fase va constar de diverses reunions i intercanvis de correus electrònics amb l'ICS on es van fixar els objectius i el pla del projecte. S'han previst algunes reunions més per acabar de fixar detalls per això aquest fase s'allarga fins al març. Aquesta tasca es solapa tant amb GEP com amb els anàlisis manuals de les dades però és una tasca amb una càrrega de treball baixa.

Va durar des de principis de gener fins a mitjans de març.

Preprocessament de les dades, anàlisis manuals i tria de mètodes a usar

Aquesta fase va començar tant bon punt es disposa de les dades de prova. Primerament es va estudiar l'estructura i contingut de les dades per entendre-les i se'n va fer una comprovació d'integritat.

Les anàlisis manuals es van portar a terme sobre les taules de problemes de salut i receptes de medicaments. Després d'aquests es van fer anàlisis conjuntes de les dues taules.

Va començar al 21 de gener i va durar fins a mitjans d'abril.

Automatització, construcció del software

Es va automatitzar l'extracció i filtratge de les dades, els experiments i es va crear una interfície que pugués mostrar els resultats de manera comprensible per als experts. Aquesta fase va consistir en implementar els mètodes seleccionats en la fase anterior ajuntant-los en un mateix programa. Tenia una càrrega de treball alta però previsible.

Va començar a finals d'abril i va acabar a finals de maig.

Experimentació i validació dels resultats

Els resultats són validats estadísticament i els experts de l'ICS els valoren. En aquesta fase s'hauran de fer reunions amb els responsables i experts de l'ICS. Aquesta fase s'allargarà durant l'estiu ja que el programa produeix molta informació que els experts hauran d'analitzar en deteniment.

Fase final

En aquesta fase final es va redactar un manual d'usuari, es va acabar la memòria del projecte i es va preparar la presentació oral.

Va començar la última setmana de maig i va acabar a mig juny.

4.1 Pressupost

Recursos humans

Aquest apartat es fa suposant que l'autor del projecte cobrés com un professional treballant per a una empresa, és a dir, comptant quotes patronals, part corresponent de despeses fixes,etc. En canvi tant en els experts del ICS com en els directors del projecte no s'hi afegeix el cost extra de treballar per una empresa donat que la seva participació en el projecte no afecta les seves nòmines.

Rol	Temps de dedicació	Preu/hora	Cost total
Gestor de projectes	120 hores	50 €/hora	6.000,00 €
Data scientist	250 hores	60 €/hora	15.000,00 €
Enginyer de software	80 hores	30 €/hora	2.400,00 €
Director del projecte	20 hores	30 €/hora	600,00 €
Codirectora del projecte	10 hores	30 €/hora	300,00 €
Metges ICS	20 hores	30 €/hora	600,00 €
Informàtics ICS	15 hores	30 €/hora	450,00 €
Total estimat	**515 hores**		**25.350,00** €

D'aquestes 515 hores 450 són meves. 75 hores de gestor de projectes corresponen a GEP (3 crèdits ECTS x 25 hores/crèdit) i les altres 375 als 15 crèdits del treball final.

Hardware

Recurs	Preu	Vida útil	Cost d'amortització (6 mesos)
Mountain Performance 15 (portàtil)	975 €	5 anys	97,50€

Software

El pressupost destinat a software serà de **0 €** ja que tot el software que s'utilitzarà té llicencia gratuïta exceptuant una llicencia de Windows 7, a la que no s'imputa cost perquè ja està amortitzada al tenir més de tres anys.

Pressupost total

Concepte	Cost
Recursos humans	25,350,00 €
Hardware	97,50 €
Software	0,00 €
Total	**25.447,50 €**

Els costos s'han adaptat +/- 5% a la planificació prevista.

5 PRELIMINARS

5.1 Dades disponibles

Per elaborar el treball s'ha disposat de les dades corresponents als centres de l'ICS de l'àrea metropolitana de Barcelona de l'any 2013. Les vam rebre de l'ICS el 24 de gener.

Les dades corresponen principalment a

• Les taules de farmàcia, és a dir, taules que indiquen què s'ha receptat a cada pacient, qui li ha receptat i quan li han receptat, i

• Les taules de problemes de salut, que indiquen quins problemes de salut té cada pacient. Aquesta última taula inclou també aquells problemes de salut actius previs a l'any 2013 com podrien ser malalties cròniques.

Addicionalment hi ha diverses taules més petites que proporcionen informació extra com la de població assignada que indica edat i sexe dels pacients i els catàlegs de medicaments i malalties.

Per tal de complir amb la regulació vigent de protecció de dades s'ha firmat un contracte de confidencialitat i les dades han estat anonimitzades: prèviament a la cessió de les dades, l'ICS va aplicar el seu algorisme estàndard d'anonimització que va traduir el codi de la seguretat social de cada pacient (CIP) a un identificador aparentment aleatori, però evidentment mantingut a través de les diverses taules per poder usar-lo com a clau de pacient.

Neteja de les dades

A causa de l'algorisme d'encriptació (l'algorisme estàndard d'anonimització de dades que s'utilitza a l'ICS) s'ha hagut de descartar totes aquelles dades de pacients que tenien un CIP provisional, ja que l'algorisme no funcionava en aquests casos i els traduïa a NULL cosa que feia que no es poguessin creuar les taules correctament.

També s'han descartat les dades de pacients que no tenien una entrada a la taula de pacients assignats, és a dir, que apareixien a les taules de prescripcions o problemes de salut però no es podia saber l'edat o el sexe a causa de que no apareixia a la taula de pacients assignats.

S'ha treballat només amb aquells problemes de salut i prescripcions actives a data de 1 de gener de 2014, però seria senzill utilitzar dades d'altres dades per poder comparar-les.

Per utilitzar les dades de prescripcions només s'ha tingut en compte el codi del principi actiu. Els principis actius es classifiquen amb els codis ATC[2]. Els codis ATC es componen de fins a 7 caràcters alfanumèrics. Podem agregar fàcilment els codis ATC eliminant darrers caràcters de cada cadena. Com més caràcters s'eliminen més agregació es té. Per exemple, l'Omeprazol correspon al codi A02BC01 i el codi A02BC correspon als inhibidors de la bomba de protons que engloba 4 principis actius similars més. Implícitament quan s'utilitza la classificació per principi actiu s'estan descartant totes aquelles receptes que no contenen principis actius com poden ser les gases. Tampoc es fa distinció entre diferents marques de medicaments que subministren el mateix principi actiu.

Els codis de malalties utilitzen la Classificació Internacional de Malalties (CIM-10)[3]. Els codis són compostos per una lletra dos dígits i opcionalment un punt amb un o dos dígits (per exemple, A79.1). Són més difícils d'agregar que els codis de medicaments, perquè alguns grups poden anar, per exemple, de A50 a A64 de manera que es requereix introduir la classificació manualment. En aquest treball quan s'han agregat els medicaments només s'han retirat els dígits de després del punt.

Al final per a cada pacient, es guarda el seu identificador, el sexe, la edat, la àrea bàsica de residència i la llista de malalties i medicaments que pren.

Quantitat de dades

En total, les dades són de 1,6 milions de persones amb un total de 12,3 milions de problemes de salut i 7,7 milions de prescripcions. D'aquestes aproximadament un 15% corresponen a usuaris amb els CIP (identificador) NULL o no apareixen a la taula de pacients assignats i per tant, no es podien creuar amb les altres taules. En total les dades sense filtrar i descomprimides ocupen 8,5 GB (560 MB comprimides), tot i que un cop s'han creuat i filtrat les dades, ocupen uns 200MB.

Limitacions de les dades

Recordem que les dades només corresponen als centres d'atenció primària del ICS i, per tant, no hi ha informació d'hospitals ni de centres privats. Això fa que no es puguin portar a terme determinats estudis geogràfics de tota la ciutat de Barcelona o que almenys s'hagi de tenir en compte el biaix que poden tenir les dades en barris amb un poder adquisitiu

2 http://es.wikipedia.org/wiki/C%C3%B3digo_ATC
3 http://ca.wikipedia.org/wiki/CIM-10

més elevat, ja que les persones d'aquests barris tindran més tendència a utilitzar la sanitat privada.

Tampoc es poden fer estudis sobre la evolució en el temps dels pacients, per exemple, ja que les dades de que es disposa només són de l'any 2013.

També s'ha de tenir en compte que les dades són entrades a la base de dades per persones (els metges) amb processos no del tot mecanitzats, i que poden ser errònies en alguns casos. Per exemple: un pacient va a l'ambulatori perquè té una grip, el metge introdueix que el pacient té la grip; quan el pacient es troba bé deixa d'anar a l'ambulatori i si el metge no pensa a posar una data de fi a la malaltia, la malaltia seguirà constant com a activa indefinidament.

5.2 Elements de Data Mining

En aquest apartat es presenten els fonaments teòrics i els algorismes en que es basen les tècniques de mineria de dades implementades. En el capítol següent explicarem per què vam triar precisament aquestes tècniques i no d'altres per abordar els problemes objecte d'aquest treball.

Frequent itemsets

El problema dels itemsets freqüents s'ha aplicat en molts dominis on les relacions entre els elements són molts a molts. El més famós és el “market-basket” que consisteix en trobar quins conjunts de productes en un supermercat són comprats freqüentment alhora per els seus clients. Per exemple, podem esperar que els clients comprin alhora llet i cereals.

Sigui $I=\{i_1, i_{2,\ldots}, i_n\}$ un conjunt de *n* atributs binaris anomenats *ítems.* Una *transacció*, també anomenada *itemset*, és un subconjunt dels items de *I* . Una *base de dades* és és un multiconjunt de transaccions. És a dir, una mateixa transacció pot aparèixer diversos cops en la base de dades.

Sigui *S* un subconjunt de *I.* Definim el suport de *S* en la base de dades *D* com el nombre o la proporció de transaccions de *D* de les que *S* n'és subconjunt. El denotem amb *suport(S,D)*, o bé *suport(S)* si *D* se sobreentén. Quan el suport és expressat en termes relatius també es pot anomenar freqüència. En la implementació, quan parlem de suport estem parlant sempre en termes relatius (freqüència).

Anomenem *itemsets freqüents* aquells subconjunts $F \subseteq I$ tals que *suport(F)* >= σ on *σ* és el suport mínim, que pot estar expressat tant en termes relatius com en valors absoluts i que serà sempre clar en el context.

Anomenem **Itemset Mining** a la tasca de trobar tots els itemsets freqüents de *D*. Aquesta tasca pot suposar un repte perquè l'espai de cerca potencial és exponencial en el nombre d'items. Concretament hi ha $2^{|I|}$ ítem sets diferents.[10] [11]

L'espai de cerca efectiu es pot reduir gràcies a la següent propietat, aplicable ja que ens interessen només els itemsets freqüents:

Proposició: Support monotonicity

Siguin $X, Y \subseteq I$ dos ítem sets. Llavors:

$$X \subseteq Y \Rightarrow suport(Y) \leqslant suport(X)$$

Això significa que si un ítem set de mida *n* és freqüent llavors tots els seus subconjunts de mida *n-1* o menys han de ser freqüents també. S'aplica algorísmicament en el sentit contrari: en el moment que sabem que un conjunt no és freqüent, podem deixar de considerar tots els seus superconjunts. Començarem explorant els conjunts per ordre de mida (0, 1, 2, ...) per descartar al més aviat possible el màxim de candidats.

Hi ha diversos algorismes que resolen el problema del Frequent Itemset Mining de manera relativament satisfactòria, com poden ser Apriori, Eclat o FPGrowth [10]. Tots ells utilitzen d'alguna manera o altra la propietat de Support Monotonicity.

Paral·lelització d'Itemset Mining

Tot i tenir bons algorismes i bones implementacions podem acabar tenint un problema de rendiment a mesura que creixi el nombre de transaccions, suposant que el nombre de ítems no creixi també. És important que els algorismes que utilitzem puguin escalar horitzontalment per tal de poder mantenir els costos de computació lineals amb les dades que estem tractant.

Encara que en l'àmbit d'aquest treball no hem arribat a aplicar paral·lelisme, sí que hem investigat lleugerament la possibilitat d'aplicar-lo a aquesta tasca. L'algorisme de Savasere, Omiecinski i Navathe (SON) [11] consisteix en particionar *D* en *p* parts fent que cada transacció vagi a parar aleatòriament a una part i que les parts tinguin la mateixa mida. Executem un algorisme de *Itemsets mining* a cada partició però utilitzant com a suport mínim s/p on s seria el suport mínim original.

Un cop hem processat totes les parts podem agafar la unió dels ítem sets candidats que ens ha donat cada part. Cal veure que si un item set no és freqüent a cap part llavors el seu suport serà menor que s/p, és a dir que no ens hem deixat cap ítem set freqüent en aquest pas (no hi ha falsos negatius).

Fem una altra passada per veure si els ítem sets candidats tenen el suport mínim s. D'aquesta manera descartem els possibles falsos positius que haguessin pogut aparèixer al primer pas.

Regles d'associació

Una regla d'associació és una expressió del tipus $X \Rightarrow Y$ on X i Y són itemsets tals que $X \cap Y = \emptyset$. Aquesta regla expressa que si una transacció conté tots els ítems de X, llavors la transacció també conté tots els ítems de Y. X s'anomena *body* o *antecedent* i Y s'anomena *head* o *conseqüent*.

El suport de $X \Rightarrow Y$ a D és el suport de $X \cup Y$.

La confiança d'una regla és la probabilitat condicionada de tenir Y donat que tenim X a la transacció:

$$confiança(X \Rightarrow Y, D) := P(Y|X) = \frac{suport(X \cup Y, D)}{suport(X, D)}$$

La regla s'anomena *confiable* si *P(Y | X)* supera un llindar preestablert mínim *conf* entre 0 i 1. El valor de *conf* serà clar sempre del context. Notem que les implicacions de la lògica clàssica són les corresponents a *conf* = 1.

Un altre cop tenim diversos algorismes que resolen el problema de trobar regles d'associació. La majoria d'ells, com ara Apriori, procedeixen primer trobant tots els frequent itemsets, i després usant-los per construir totes les possibles regles amb suficient confiança.

Odds ratio

Serveix per quantificar com de forta és l'associació entre a i b. En aquest cas com més alt és el odds ratio més alta és l'associació.

Sigui a = 1 si a apareix a la transacció i 0 en cas contrari i b = 1 si b apareix a la transacció i 0 en cas contrari. p_{xy} Ens indica si x o y apareix a la transacció.

$$Odds(a,b) = \frac{p_{11} * p_{00}}{p_{10} * p_{01}}$$

P-values

El test d'independència entre dues variables Chi-Square pot ser aplicat quan tenim dues variables A i B categòriques d'una sola població. La hipòtesis nul·la d'aquest test diu que conèixer el valor de la variable A no ajuda a conèixer el valor de la variable B, és a dir, A i B són independents.

Primer cal conèixer els **graus de llibertat** (DF) que es calcula com:

$$DF = (r-1)\cdot(c-1)$$

On r i c serien el nombre de valors diferents que poden pendre les variables A i B respectivament.

Seguidament calculem les **freqüències esperades**:

$$E_{r,c} = \frac{(n_r \cdot n_c)}{n}$$

on $E_{r,c}$ és la freqüència esperada per un parell de valors de A i B. $n_r n_c$ Són el nombre d'observacions de la variable A amb el valor r i la variable B amb el valor c. n és la mida total de la mostra.

El **test estadístic** és una variable aleatòria chi-square (X^2) definida per:

$$X^2 = \frac{\sum (O_{r,c} - E_{r,c})^2}{E_{r,c}}$$

on $E_{r,c}$ és la freqüència esperada per el valor r de la variable A i el valor c de la variable B i $O_{r,c}$ és la freqüència observada per el valor r de la variable A i el valor c de la variable B.

P-value: Un cop tenim l'estadístic, podem transformar-lo en la probabilitat que A i B estiguin relacionades utilitzant una taula de la distribució Chi-square utilitzant els graus de llibertat calculats al primer pas.

Per més informació es pot consultar [12].

Pointwise Mutual Information (PMI)

Definim:

$$PMI(a,b) = \log\left(\frac{p(a,b)}{p(a)\cdot p(b)}\right)$$

On p(a,b) és la probabilitat de trobar les malalties a i b en un pacient, p(a) és la probabilitat de trobar a en un pacient i el mateix en b.

És a dir si a, b no tenen relació entre elles PMI(a,b) és proper a 0. Si a i b tenen tendència a aparèixer juntes PMI(a,b) és superior a 0 i, si PMI(a,b) és inferior a 0 significa que a i b tendeixen a no aparèixer juntes.

Podem acotar el resultat normalitzant el PMI entre [-1,1] en comptes de $(-\infty,\min(-\log(p(a)),-\log(p(b)))]$. Això significaria que si a i b es troben sempre junts tindrem un NPMI(a,b) = 1 i si mai es troben junts NPMI(a,b) = -1

$$NPMI(a,b) = \frac{PMI(a,b)}{-\log(p(a,b))}$$

Per més informació consultar [13].

6 ALGORISMES I TÈCNIQUES DE MIRERIA DE DADES

En aquest apartat s'explica en detall com s'han implementat cada una de les funcionalitats descrites en l'apartat 3 i que constitueixen el cor de l'aplicació.

6.1 Generació de grafs de malalties i medicaments

En aquest apartat només es parlarà de malalties, però tot el que s'explica s'aplica també en el cas de els medicaments.

Volem generar grafs de malalties relacionades entre elles, de manera que cada malaltia sigui un node i cada aresta representi una relació. Diem que una malaltia té relació amb una altre si la parella apareix més del que seria esperat en els pacients. Posarem un pes a cada aresta del graf de manera que l'usuari podrà elegir en el moment de consultar-lo quin és l'interval de pes que vol que es mostri. Si posa un pes més baix es mostraran arestes menys relacionades i si posa un pes més alt es mostraran les arestes més estretament relacionades.

Per simplificar els càlculs aprofitarem l'algorisme de freqüent itemsets per comptar aquells itemsets de com a molt mida 2. D'aquesta manera ja tenim les probabilitats que necessitem per fer els càlculs que venen a continuació i es simplifica el cost computacional i de programació.

Per quantificar el pes s'han estudiat tres opcions diferents:

Odds ratio

Serveix, per quantificar com de forta és l'associació entre a i b. En aquest cas com més alt és el odds ratio més alta és l'associació. Les malalties que tendeixin més a aparèixer juntes tindran un odds més alt que les que no.[7]

P-values

En aquest cas ens donarà un p-value més proper a 1 com més confiança hi hagi entre l'associació entre a i b. Es quantifica la relació entre les dues variables sense tenir en compte si tendeixen a estar juntes o separades. A més, com més casos en termes absoluts tinguem de A o B, més proper a 1 serà el p-value. Per això podem esperar que associacions entre la hipertensió i la obesitat estiguin entre aquelles amb un p-value més

alt. Aquestes associacions com que es deuen a malalties freqüents entre la població són àmpliament conegudes i no interessa tant trobar-les segurament.[7]

Pointwise Mutual Information (PMI)

Aquesta mesura s'ha utilitzat en estudis que han analitzat la relació entre ingredients de les receptes que pengen els usuaris en una pàgina web [14] però no es coneix de cap cas que s'hagi utilitzat en les relacions entre malalties o medicaments

Si A i B tendeixen a aparèixer juntes el PMI serà alt si en canvi si A i B tendeixen a estar separades el PMI que tindran serà negatiu, mentre que si no tenen relació serà proper a 0.

Aquesta és la solució implementada en l'aplicació final, ja que ens permet distingir fàcilment entre aquelles relacions que fan que A i B apareguin juntes, com aquelles que fa que apareguin separades. Això pot ser molt útil si volem estudiar aquelles malalties o medicaments incompatibles.

En la implementació s'han utilitzat logaritmes en base 10, si trobem per exemple, una relació PMI(a,b) = 2 sabrem que la probabilitat de trobar a i b juntes, és 10 vegades més del que s'esperaria.

6.2 Generació de regles d'associació

L'objectiu d'aquest procés és trobar regles que ens diguin quins medicaments prenen els pacients amb una determinada malaltia, de manera que només ens interessen aquelles regles que van de malalties a medicaments o de medicaments a malalties.

Per a aquesta tasca, ens vam adonar aviat que considerar grafs com els de l'apartat anterior no és satisfactori. Si considerem grafs on cada node és una malaltia o un medicament, hi ha relacions n-àries interessants (amb n>2) que no poden inferir-se del graf únicament. Per exemple, considerem tres malalties x, y i z, i suposem que calculem l'associació entre cada parell d'elles, per posar-lo en un graf. Segurament hi ha combinacions de valors que no ens permeten distingir entre els dos següents casos:

- Molts pacients pateixen x i y (però no z), o bé x i z (però no y), o bé y i z (però no x)
- Molts pacients pateixen alhora x, y i z.

Això ens portaria a considerar estructures més complicades que aquests grafs, per capturar aquestes relacions n-àries. Podríem considerar grafs on els nodes són conjunts de malalties i/o medicaments. O bé podríem considerar *hipergrafs,* on cada node és

encara una malaltia o medicament, però les *hiperarestes* són conjunts de nodes de cardinalitat potser més gran que 2 – essent les arestes tradicionals en grafs les de cardinalitat 2.

Resulta que llavors el problema de calcular totes les hiperarestes presents (amb un pes determinat) en un conjunt de dades que implícitament representa un hipergraf és, formalment, el mateix problema que calcular tots els frequent itemsets en el conjunt de dades. De fet, el problema de trobar tots els itemsets d'ha anomenat de vegades (però menys sovint) el problema de *hypergraph transversal* [15]. És per això que decidim abordar aquest problema en terminologia d'itemsets i regles d'associació. Això no vol dir que no es pugui extreure informació molt útil del graf bàsic amb tècniques, per exemple de xarxes complexes (cerca de comunitats, influenciadors, mesures de centralitat, etc.) però degut al temps disponible es deixa per a possibles ampliacions futures.

El primer pas per generar les regles consisteix en trobar els itemsets freqüents utilitzant un algorisme com podria ser Apriori o Eclat [10]. No s'utilitzarà Apriori per generar totes les regles (a la implementació final, sí que s'utilitza Apriori per obtenir els itemsets freqüents) perquè no és necessari, ja que només volem un subconjunt molt concret d'aquestes. Per tant, hem fet una implementació pròpia de la manera que s'explica a continuació.

Un cop tenim tots els itemsets freqüents elegim aquells que contenen malalties i medicaments alhora i calculem la confiança de malalties a medicaments utilitzant el suport calculat als altres itemsets descartats anteriorment.

Per exemple si A, B i C són malalties i X Y Z medicaments i S_1 el suport de {A, B, C} i S_2 el suport de {A, B, C, X, Y, Z}:

La regla {A, B, C} => {X, Y Z} tindrà confiança $\frac{S_2}{S_1}$.

A partir de les reunions amb els experts de l'ICS es va determinar que el que fa que un metge doni un determinat medicament és una sola malaltia, per tant, les regles podien ser 1 a 1. Això simplifica molt la complexitat d'eliminar regles poc interessants.

Descartar regles poc interessants

En aquest apartat es descriuen dos mètodes per descartar regles i ítem sets que no ens interessen.

L'objectiu final d'aquest procediment és obtenir regles que permetin saber quins medicaments es donen a conseqüència d'una malaltia. Però les regles només ens

indiquen que als pacients que tenen A se'ls dona X amb una certa confiança. Podem esperar que apareguin regles A => X però realment un metge no dona X a conseqüència de A sinó perquè A apareix freqüentment associat amb una altre malaltia B que fa que el metge li recepti X al pacient. Per tant, tindrem una regla B => X que tindrà una confiança superior a la confiança de A => X. És a dir, que la regla A => X no ens interessa.

Mètode 1:

Un cop tenim les regles descartem aquelles que han aparegut a causa de l'aleatorietat. Per exemple, A => B si $\frac{p(A,B)}{p(A)\cdot p(B)} \simeq 1$ podem descartar la regla. Més en general, si tenim un ítem set freqüent S on A i B són dos subconjunts disjunts de S. Si $\frac{p(A,B)}{p(A)\cdot p(B)} \simeq 1$ podem descartar aquell ítem set perquè per la S*upport monotonicity* existirà un altre item set freqüent S' que serà subconjunt de S i ens aportarà la mateixa informació que S però serà més petit i per tant, més comprensible.

Mètode 2:

Podem descartar una regla A => X amb confiança C_1 si generant les regles de malalties a malalties ens trobem que existeixen dues regles B =>X amb confiança C_2 i B => A amb confiança C3. Tals que $C_1 \simeq C_2 \cdot C_3$.

S'ha observat que aquest mètode descarta correctament regles que vinculen una malaltia poc freqüent amb un medicament que es dona per a malalties molt freqüents com pot ser la insulina per la diabetis.

6.3 Detecció d'episodis no tancats i medicaments sense explicació

Utilitzant les regles de l'apartat anterior podem detectar pacients que no compleixen el que s'esperaria. És a dir, que no podem explicar les malalties que tenen per els medicaments que tenen receptats o a l'inrevés.

Segons si intentem explicar els medicaments a partir de les malalties o a l'inrevés podem detectar coses diferents.

En el cas que intentem trobar un medicament que expliqui una malaltia, el que volem trobar són el que s'anomena episodis no tancats: quan un pacient va al metge per una malaltia puntual i aquest li recepta alguna cosa, les receptes tenen una data de caducitat però les malalties no; llavors quan el pacient passa a trobar-se bé i no torna el metge, la

malaltia queda oberta indefinidament a la base de dades. Per exemple, a un pacient que va patir un accident de moto li consta una fractura de braç com a problema de salut obert des de fa mes de dos anys. Cal notar però, que aquest mètode pot donar molts falsos positius sobretot si no es filtren correctament malalties cròniques per a les quals no es recepta res a vegades quan la malaltia és lleu, com pot ser la hipertensió.

En el cas que intentem trobar una malaltia que expliqui un medicament el que volem trobar són medicaments que no s'expliquen donat les malalties que té el pacient. En aquests casos s'hauria d'investigar si es deu a que el metge no ha introduït la malaltia a la base de dades o alguna altre cosa.

Per detectar aquests casos, per a cada pacient, s'ha de mirar si existeix algun subconjunt de les seves malalties que encaixi amb l'antecedent d'alguna regla, en aquest cas mirem si el conseqüent és subconjunt del conjunt de medicaments del pacient. Si és així, diem que les malalties de l'antecedent tenen explicació, en cas que al final quedi alguna malaltia sense explicació etiquetem el pacient com a sospitós.

Podem veure que si per a cada pacient, hem de generar tots els subconjunts de malalties, això pot resultar un problema greu ja que el nombre de subconjunts creix de manera molt ràpida amb el nombre total d'elements del conjunt i hi ha pacients amb més de 50 malalties. Per sort, com que les regles només són 1 a 1 només hem de generar subconjunts de mida 1. És a dir el cost és lineal amb el nombre d'elements.

7 IMPLEMENTACIÓ

El programa és bàsicament un servidor senzill que permet als usuaris connectar-se des d'una interfície web. Aquest programa engloba totes les funcionalitats descrites en l'apartat anterior i les presenta de manera que els usuaris puguin consultar-ne els resultats tractar amb les dades i executar nous experiments de manera fàcil.

Aquesta implementació, és un prototip per tal de posar a prova les tècniques descrites en l'apartat anterior, per tant, manca de característiques que s'haurien d'esperar d'una aplicació final, com pot ser la seguretat d'accés a les dades i més eines per treballar i modificar les dades o la capacitat de guardar i repetir experiments.

Tecnologies usades

Per elegir les tecnologies per utilitzar s'ha buscat que fossin el màxim de portables i que es pogués programar ràpidament un prototip amb aquestes. Per això s'ha elegit Java per el llenguatge *back-end* i JavaScript per a la interfície i el servidor.

El software té tres parts molt diferenciades. El servidor, el mòdul d'anàlisi i la interfície. La comunicació entre les tres parts es codifica amb JSON[4] que és un format per a l'intercanvi de dades. JSON és interpretat de forma nativa per JavaScript i per Java s'utilitzarà una petita llibreria[5] per interpretar-lo.

7.1 Servidor

El servidor és un procés que està escoltant un port, quan un usuari es connecta li demana que introdueixi el nom d'usuari. Quan l'usuari introdueix el seu nom, es crea un procés en Java que executa un Jar. Des d'aquest moment el servidor és un simple router que només s'encarrega de redirigir les comandes que envia la interfície cap al procés Java que les interpreta i retorna una resposta.

Pot executar-se tant en mode local, com en un servidor connectat a Internet. En aquest segon cas diversos usuaris poden connectar-se al mateix servidor i executar-hi experiments i fins i tot compartir-ne els resultats entre ells. Si el servidor fos proper a la base de dades podria actualitzar-se de forma automàtica extraient-ne les dades i netejant-les per tal de poder ser utilitzades.

4 http://json.org/
5 http://www.json.org/java/index.html

Està implementat completament amb Node.js[6] (JavaScript) i és un programa senzill de no més de 150 línies de codi.

7.2 Mòdul d'anàlisi

El mòdul d'anàlisi és un procés Java que s'encarrega d'executar les comandes que rep per la entrada estàndard i el resultat també l'escriu per el canal estàndard de sortida. Això fa que es pugui executar com una aplicació normal per línia de comandes i sigui completament independent de la resta del programa.

Tant les instruccions que llegeix com les instruccions que escriu són en format JSON. Això fa que els resultats siguin fàcilment interpretables per la interfície que està implementada en JavaScript i es puguin expressar instruccions bastant complexes.

El procés interactua amb fitxers que guarden els resultats. Hi ha tres tipus de fitxers: grafs, fitxers de regles i fitxers de pacients. Cada línia del fitxer representa una aresta del graf, una regla o un pacient respectivament. Cada columna és separada per un espai. Els tipus de dades s'identifiquen perquè cada columna va precedida per un identificador i dos punts per exemple cip:1234 addicionalment es poden afegir atributs a cada element separats per ";". El format és doncs:

nom-columna:valor;clau-attr1=val-attr1;clau-attr2=val-attr2 nom-columna2:valor2...

Els tipus d'instruccions que pot llegir el procés són bàsicament tres:

- Instruccions sobre fitxers
- Executar experiments
- Retornar resultats

Les **instruccions sobre fitxers** són instruccions que permeten descarregar fitxers o esborrar-los. A continuació s'explica en més detall les instruccions per executar experiments i retornar resultats.

Executar experiments

El procés Java permet executar experiments de forma molt flexible de manera similar a eines de data mining com ara Weka o Knime, en versió línia de comandes.

Les parts d'un experiment s'anomenen mòduls. Per poder executar un mòdul cal indicar-li l'identificador dels fitxers d'entrada i l'identificador dels fitxers de sortida, a més de les

6 http://nodejs.org/

opcions de pròpies de cada mòdul. En un experiment, es poden executar un nombre il·limitat de mòduls encara que tinguin dependències entre ells.

Hi ha molts tipus de mòduls des de filtres, fins a mòduls que poden executar elements externs com pot ser el mòdul que executa l'algorisme d'Apriori.

El programa decideix dinàmicament l'ordre en que s'executen els mòduls a partir de les dependències de fitxers que té cada un. En cas que hi hagin dead-locks s'emet un error.

Es poden estendre fàcilment les funcionalitats del programa només afegint mòduls que heretin la classe "Module".

A continuació es detalla la llista de mòduls implementats:

RowsFilter

Agafa les files d'un fitxer i les filtra segons una llista de condicions que se li han passat per paràmetre. Només rep un fitxer d'entrada però pot rebre més d'una llista de condicions, en aquest cas hi haurà més d'un fitxer de sortida i cada un contindrà les files que compleixen les condicions d'una de les llistes. Per exemple, podem dividir els pacients amb més de cinc malalties i els que tenen menys de cinc malalties.

Cada llista de condicions pot ser una llista AND o OR si és AND significa que perquè una fila s'escrigui al fitxer ha de complir totes les condicions de la llista, en el cas que sigui OR només cal complir una de les condicions de la llista. Cada condició té un tipus de columna, una condició i un valor. Per exemple, dir que l'edat ha de ser superior a 30. La possibilitat d'introduir condicions que combinin ANDs i ORs, encara que no difícil tècnicament, és una mica tediosa de fer i poc prioritària segons els membres de l'equip, i es considerarà a versions futures.

Altres coses que es poden fer es filtrar els pacients amb una malaltia o un tipus de malalties, com obtenir els pacients que tenen una malaltia relacionada amb el cor.

ColumnsFilter

Aquest filtre permet eliminar columnes, per exemple, podem eliminar totes les malalties que tenen un codi que comença per C (càncers). També pot eliminar columnes senceres que no interessen en un moment concret com podria ser el identificador o la zona de residència o fins i tot totes les malalties i tots els medicaments.

Igual que en el filtre per files, només tenim un fitxer d'entrada però podem especificar diverses llistes de filtres i tenir diverses llistes de sortida.

També té mètodes per agregar els codis de medicaments o malalties o eliminar els atributs de les columnes.

Apriori

Crea un procés per executar l'algorisme d'apriori, actualment s'executa una implementació de Christian Borgelt[7]. Prèviament es passa un filtre per eliminar els atributs de les columnes i així fer la entrada que es passa al procés extern compatible. La sortida són els ítem sets freqüents amb el seu suport expressat en percentatge.

AssociationRuleGenerator

Genera regles d'associació de malalties a medicaments i de medicaments a malalties amb una confiança mínima especificada. Com a entrada necessita un fitxer amb el suport dels antecedents de les regles (si fem regles malalties => medicaments, això són les malalties) i un fitxer amb el suport dels ítem sets.

ItemSetsGenerator

S'encarrega de creuar les dades de prescripcions, problemes de salut i població assignada. Com a entrada agafa tres fitxers amb aquestes dades ordenats per CIP. Descarta aquells pacients que apareixen als fitxers de problemes de salut o prescripcions però no al de població assignada.

Com que els fitxers estan ordenats per CIP aquesta operació de creuar les dades es pot fer en temps lineal.

GraphGenerator

Genera un graf a partir d'un fitxer amb ítem sets de fins a dos elements i amb el suport prèviament calculat. Calcula el PMI per a cada ítem set de dos elements. Cada ítem set de dos elements serà una aresta.

FrequentItemSetsFilter

Elimina aquells ítem sets de la manera que es descriu en el mètode 1 de l'apartat on es descriu com eliminar regles poc interesants.

FindSuspiciousPatients

Separa els pacients amb dos grups aquells que les malalties que tenen corresponen amb els medicaments que prenen i els que no. Addicionalment, marca aquelles malalties o medicaments que ha trobat que no tenen correspondència.

---

7 http://www.borgelt.net/apriori.html

Pot executar-se de dues maneres:

• Intentant trobar medicaments que no corresponen amb les malalties que té. Per fer això s'utilitzen regles de medicaments a malalties.

• Intentant trobar malalties que no corresponen amb els medicaments que té. En aquest cas s'utilitzen regles de malalties a medicaments.

SaveResult

S'encarrega de guardar el resultat, ja sigui un graf, un fitxer de regles o pacients de manera que sigui visible per l'usuari.

Retornar resultats

Aquest és la part que s'encarrega de respondre quan la interfície sol·licita una determinada part d'un fitxer. Els tipus de resultats que pot retornar són:

• Pacients, es pot retornar la informació de pacients provinents d'un fitxer

• Grafs, es poden fer cerques en grafs i retornar el subgraf generat a partir de fer una cerca en amplada a una certa profunditat, amb un pes mínim i un pes màxim a partir d'un node.

• Regles d'associació, permet buscar les regles relacionades amb un medicament o una malaltia o un conjunt d'aquests.

7.3 Interfície

La interfície és una capa que permet interactuar als usuaris amb el procés Java sense haver-se de preocupar del seu funcionament. A més, com que és una interfície web es pot fer de forma remota.

Està programada completament en JavaScript, s'ha utilitzat la llibreria jQuery[8] per a la manipulació del DOM (Document Object Model), és a dir, els elements visibles. La llibreria DataTables[9] per generar taules fàcilment i la llibreria soket.io[10] per a la comunicació client-servidor.

El primer que es demana a l'usuari quan obre la url corresponent, és que introdueixi el nom d'usuari. Quan aquest introdueix el seu nom s'estableix una comunicació client-servidor mitjançant un web-socket, de manera que la connexió es manté oberta fins

8 http://jquery.com/
9 https://datatables.net/
10 http://socket.io/

que l'usuari tanca el navegador o abandona la pàgina. Es va decidir que era millor fer-ho amb un web-socket perquè si es feia amb AJAX (peticions asíncrones) podia passar que si la resposta del servidor tardava a arribar, com passa quan s'està executant un experiment, la connexió es tanqués automàticament a causa del navegador.

La interfície principal es divideix en dues parts, com es pot veure a la Figura 2: A la part esquerra es mostren els fitxers existents, i a la part dreta les accions que podem fer amb aquests fitxers.

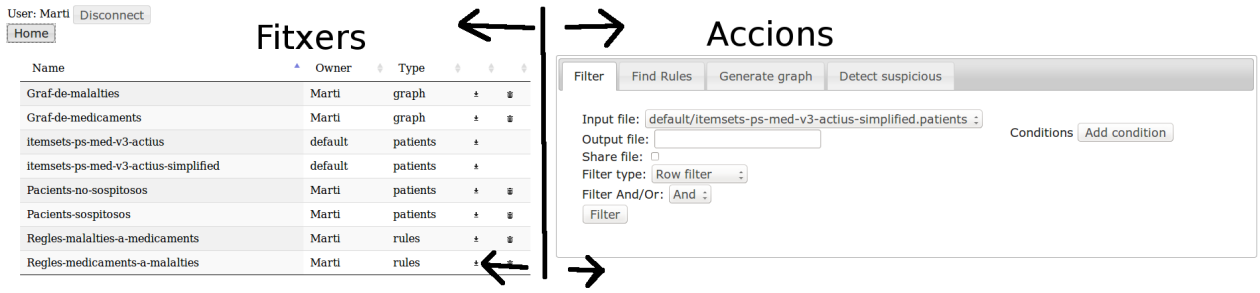


Figura 2. Menú principal de la interfície

Fitxers

Inicialment el programa només conté tres fitxers que són itemsets-ps-med-active, itemsets-ps-med-active-simplified i itemsets-ps-med-active-simplified-synonyms que contenen totes les malalties i medicaments actius el dia 1/1/2014 dels pacients dels centres d'atenció primària de l'ICS a Barcelona. El segon fitxer conté els medicaments i medecines simplificats, és a dir, ignorant les dues últimes xifres dels codis ATC i les xifres després del punt dels codis de malalties. Finalment, el tercer fitxer, conté el mateix que el segon però tractant determinats grups de malalties com poden ser les depressions o les hipertensions com si fossin la mateixa.

Hi ha tres tipus de fitxers: regles, pacients i grafs.

Cada fitxer pot ser descarregat en format CSV per a un full de càlcul, per exemple, o en cas que sigui un graf, en format GML per poder-lo consultar amb una eina de grafs. A la Figura 3 es mostra com es poden descarregar els fitxers i esborrar-los.

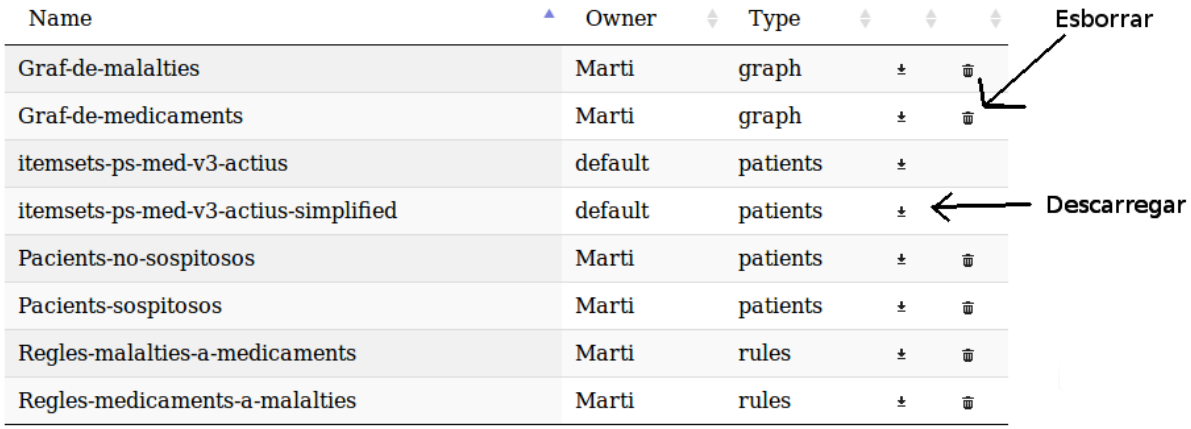


Figura 3. Llista de fitxers disponibles en una execució.

Cada fitxer té un propietari que és el qui ha creat el fitxer, només ell el pot editar, per fer que el fitxer sigui visible per als altres usuaris del sistema, ha de marcar la opció “share file”. Aquesta opció només funciona si s'està treballant en un servidor i no en un localhost.

Fitxers de regles

Els fitxers de regles mostren les regles de associació trobades a partir d'un fitxer de pacients.

Podem demanar a un fitxer de regles que ens mostri una malaltia concreta o un conjunt de malalties posant el codi al buscador. Per exemple posar I (la lletra i) al buscador ens retornaria totes les regles relacionades amb les malalties del cor.

També podem afegir regles manualment.

Grafs

Representen un subgraf d'un fitxer de grafs. Els nodes dels grafs mostren el codi de les malalties o medicaments i tenen un color diferent en funció del grup al qual pertany la malaltia: malalties del cor en vermell, càncers en marró etc. La mida dels nodes creix en funció del nombre d'arestes que té. Les arestes també són més gruixudes en funció del seu pes. La Figura 4 ens mostra una representació de la interfície d'usuari per als grafs.

Si volem conèixer el nom d'un node només cal clicar-hi a sobre i apareixerà una taula al costat indicant el nom del node i el dels seus veïns més propers, així com també el pes de les arestes per arribar als veïns.

Per consultar el graf hi ha tres paràmetres:

- El node inicial
- La profunditat màxima
- El pes mínim
- El pes màxim

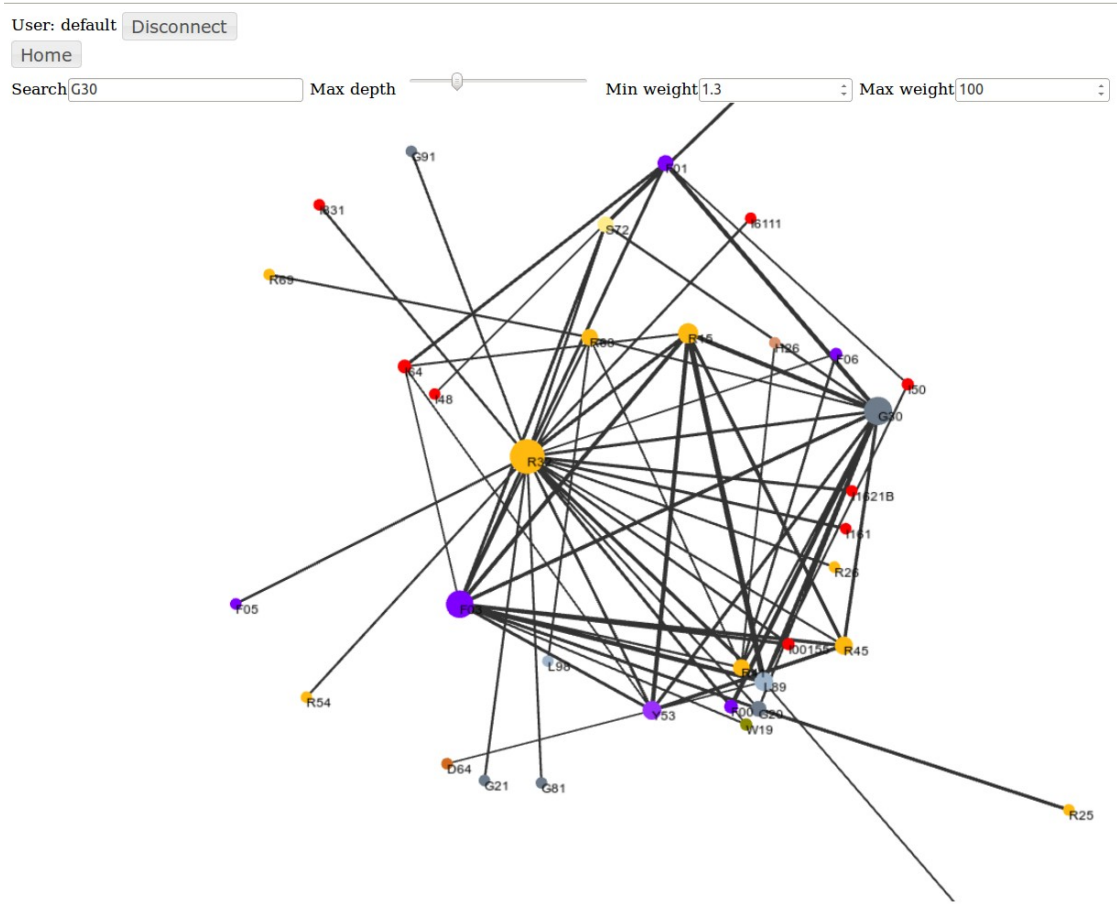
El pes de les arestes del graf ens diu quantes vegades més apareix la relació del que s'esperaria si no hi hagués una relació, és a dir que les dues variables que representen les malalties fossin independents. Cal notar que el pes de les arestes del graf és en escala logarítmica per tant si el pes entre dues malalties és 2 significa que la probabilitat de trobar-les juntes és 100 vegades més del que s'espera.

Si interessa, també es pot definir un interval negatiu. D'aquesta manera podem veure relacions que apareixen menys del que s'esperaria per atzar. Un pes de -1 ens indica que la relació apareix 10 cops menys del que s'esperaria si les ocurrències dels nodes dels extrems fossin independents.

Per dibuixar els grafs s'utilitza una llibreria de grafs en JavaScript anomenada ngraph[11], aquesta llibreria utilitza un algorisme de forces per tal de col·locar els nodes del graf en una disposició que permeti visualitzar-los a tots. Això fa que quan es carrega un graf, els nodes es vagin movent sols fins que s'estabilitzen. Pot passar que si hi ha massa nodes i massa arestes l'algorisme de forces no s'arribi a estabilitzar i es quedin movent-se indefinidament.

11 https://github.com/anvaka/ngraph

Figura 4. Representació d'un graf de malalties relacionades centrat en la malaltia de l'Alzheimer.



MALALTIA D'ALZHEIMER (G30)

relation	weight
F00 DEMÈNCIA EN LA MALALTIA D'ALZHEIMER (G30.-+)	2.631
R15 INCONTINÈNCIA FECAL	2.547
F01 DEMÈNCIA VASCULAR	2.545
L89 ÚLCERA DE DECÚBIT	2.445
F03 DEMÈNCIA, INESPECÍFICA	2.235
R45 SÍMPTOMES I SIGNES QUE INVOLUCREN L'ESTAT EMOCIONAL	2.06
Y53 EFECTES ADVERSOS D'AGENTS QUE AFECTEN PRINCIPALMENT EL SISTE	2.025
R32 INCONTINÈNCIA URINÀRIA, INESPECÍFICA	1.897
R41 ALTRES SÍMPTOMES I SIGNES QUE INVOLUCREN LA FUNCIÓ COGNOSCIT	1.754
S72 FRACTURA DEL FÈMUR	1.722
G20 MALALTIA DE PARKINSON	1.656
R68 SÍMPTOMES I SIGNES GENERALS ESPECÍFICS	1.533

Fitxers de pacients

Els fitxers de pacients mostren informació de cada pacient, tal com es mostra a la Figura 5, a part de la seva edat i sexe, també es mostren les malalties i medicaments de cada pacient, en el cas que el fitxer hagi sigut generat amb el mètode detectar aquelles malalties o medicaments que no concorden es mostrarà en vermell aquelles malalties o medicaments que no concordin.

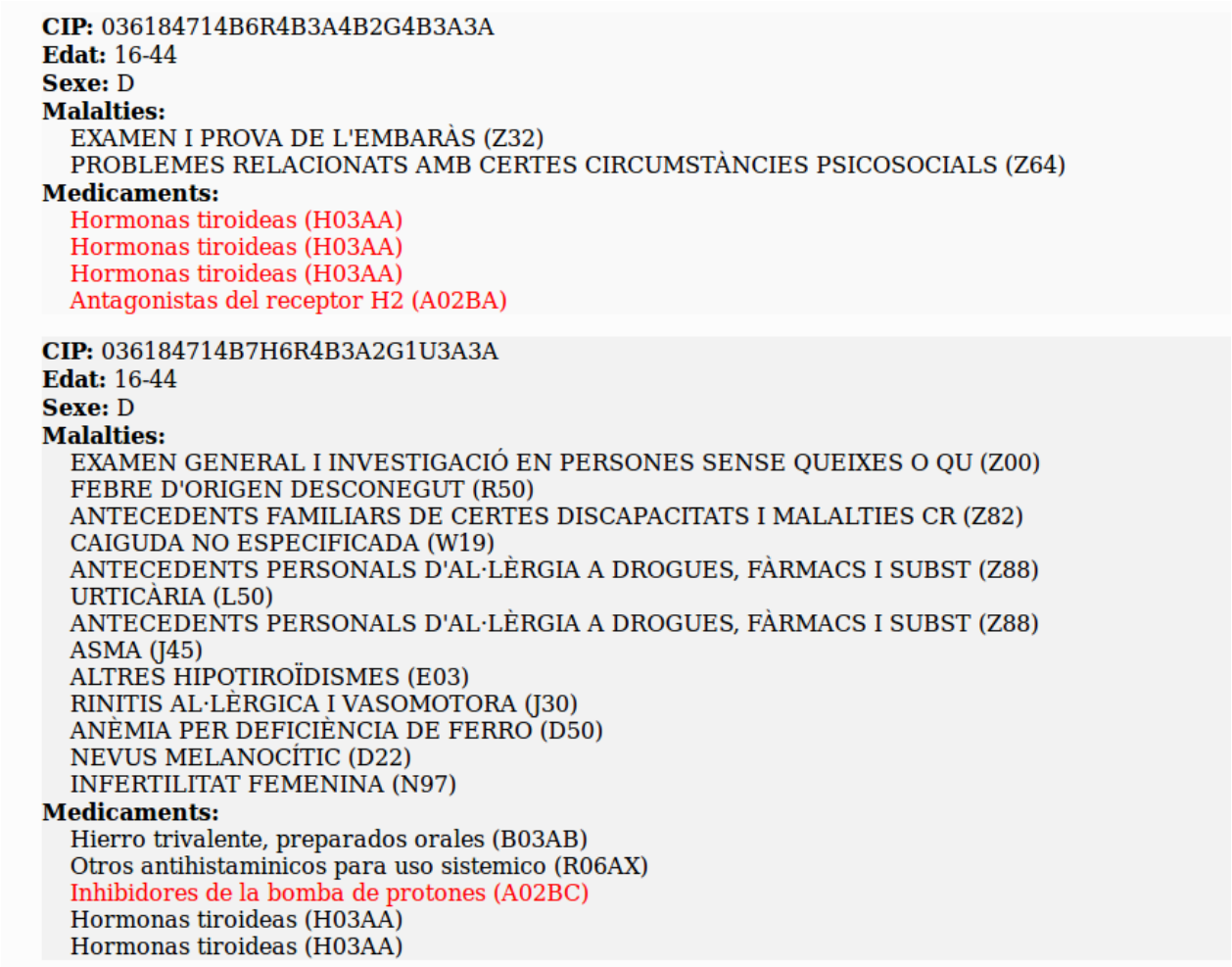


Figura 5. Mostra d'una llista de pacients que tenen medicaments que no es poden explicar per els problemes de salut que tenen.

Accions

Podem efectuar diferents tipus d'accions, cada acció agafa un o més fitxers com a entrada i treu un o més fitxers com a resultat.

Hi ha quatre tipus d'accions disponibles: Filtrar, trobar regles, generar un graf o detectar pacients “sospitosos” tal com es mostra a la Figura 6.

Filtres

Els filtres ens permeten filtrar les dades i quedar-nos amb un subconjunt de la informació del fitxer d'entrada. Hi ha dos tipus de filtres principals, filtres per files i filtres per columnes.

Els filtres per files permeten eliminar files que no compleixen un conjunt de condicions, podem filtrar per edat, sexe, malalties, medicaments, zona de residència o codi UBA. Cada condició pot ser d'un tipus diferents. Per exemple podem filtrar els pacients amb una malaltia del cor i que tinguin menys de 30 anys.

Els filtres poden ser AND o OR però, pels motius que ja s'han explicat abans, no es permeten combinacions d'aquests. Per exemple, podem seleccionar els pacients que tenen hipertensió i diabetis, o els pacients que tenen hipertensió o diabetis.

També podem seleccionar els pacients que tenen més d'un cert nombre de malalties o medicaments.

Per els filtres d'edat s'ha de tenir en compte que les dades estan entrades en forma de franges d'edat. Les franges d'edat que entrades són les que els experts de l'ICS ens van indicar com a estàndard en els seus estudis: 0-14, 15-44, 45-64, 65-74, >= 75. Per tant, en l'exemple anterior, si seleccionem els que tenen menys de 30 anys en realitat estarem seleccionant les franges 0-14 i 15-44, és a dir que hi haurà pacients que tindran més de 30 anys en el fitxer resultat. El filtre d'edat sempre intenta ser el màxim de conservador i no filtra en cas de dubte com podria ser l'interval 15-44.

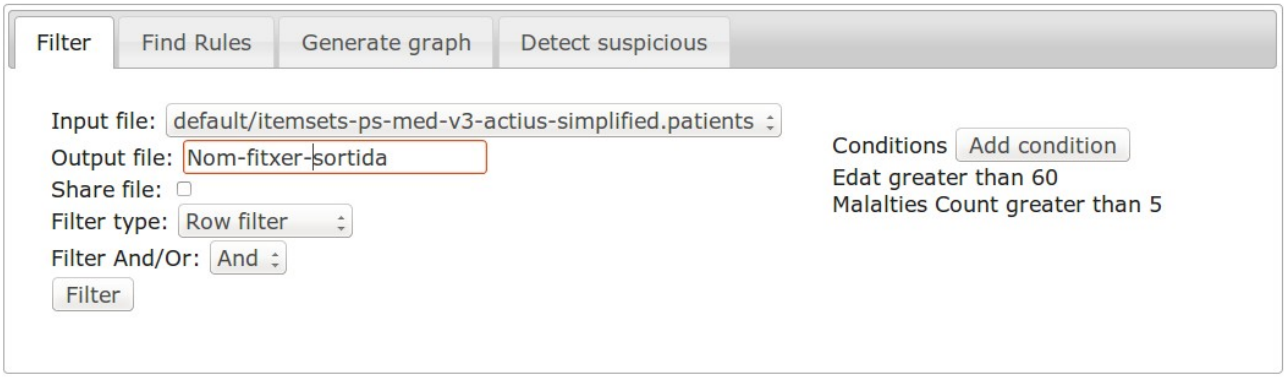


Figura 6. Apartat d'accions que es poden fer amb la interfície.

Els filtres per columnes permeten eliminar un conjunt de columnes, podem eliminar diferents columnes a la vegada. Cada columna pot ser d'un tipus diferent, per exemple podem eliminar totes aquelles malalties tals que el seu codi comença per Z i els medicaments tals que el seu codi és A02BC (Omeprazol).

Trobar regles

Aquesta acció ens permet generar un fitxer de regles a partir d'un fitxer de pacients. S'ha d'especificar un suport mínim i una confiança mínima per tal de no obtenir totes les regles possibles.

Generar un graf

A partir d'un fitxer de pacients també podem generar un graf. Es generarà un fitxer de malalties i un de medicaments. Per calcular les arestes també cal un suport mínim tal com en la part de trobar regles.

Per generar el graf és recomanable filtrar aquelles malalties i medicaments que no tinguin interès per els estudis i apareguin molt sovint com poden ser les malalties que comencen per Z o els medicaments com el Omeprazol. D'aquesta manera no apareixen tantes arestes i es pot veure millor les altres relacions.

Trobar pacients sospitosos

Aquesta acció ens permet detectar pacients que les seves malalties no concorden amb els medicaments que se'ls està donant. Per fer-ho cal especificar un fitxer de pacients i un fitxer de regles generat prèviament.

8 Experiments, resultats i avaluació

En aquesta secció es mostraran alguns dels resultats que dóna l'eina per tal de demostrar la seva correctesa i utilitat. També s'explicaran maneres de filtrar la entrada o la sortida i tunnejar els paràmetres per tal d'obtenir la informació desitjada.

8.1 Regles d'associació

Per fer aquests experiments s'han utilitzat les dades dels problemes de salut i les prescripcions actives a data de 1 de gener de 2014, agregant els problemes de salut i les prescripcions tal com s'explica a l'apartat de neteja de les dades.

Els temps d'execució, s'han calculat mesurant el temps transcorregut durant l'execució de l'experiment que consta de l'execució consecutiva dels mòduls d'Apriori, filtrar files i generar regles consecutivament.

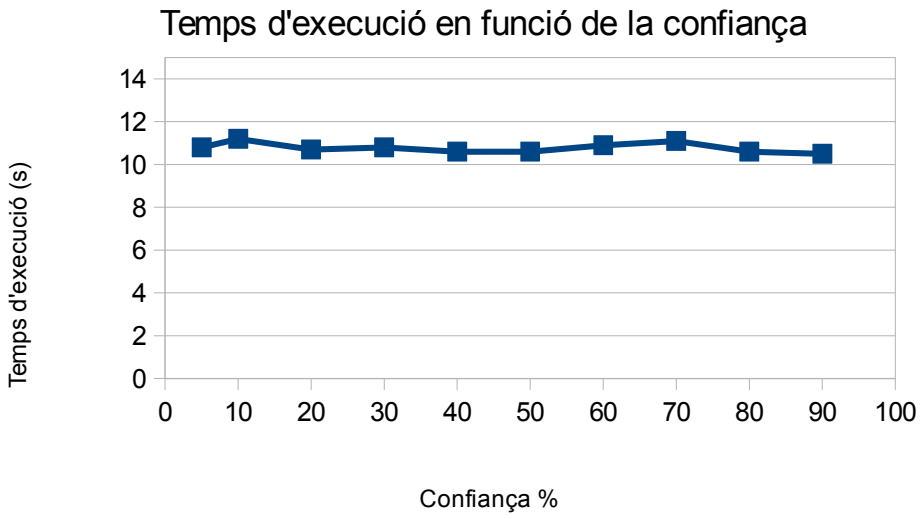


Figura 7. Temps d'execució en funció de la confiança.

La Figura 7 ens mostra que el temps d'execució del programa no varia significativament a mesura que fem variar la confiança. El suport queda fixat en un 0,05%, és a dir, que com que tenim una base de dades d'aproximadament 1,5 milions de pacients, cada regla tindrà un suport absolut de 750 casos.

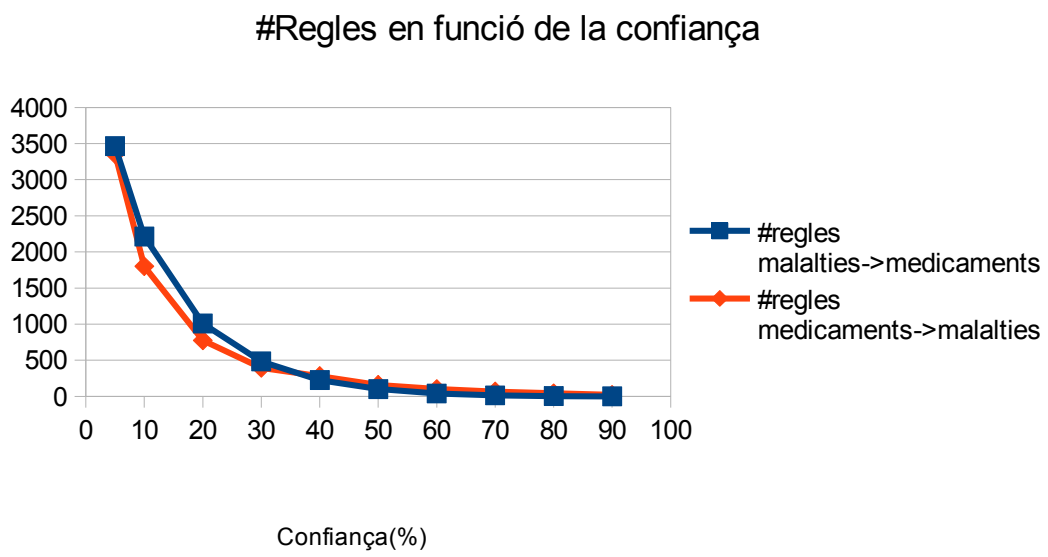


Figura 8. Nombre de regles en funció de la confiança.

La Figura 8 ens mostra com a mesura que fem augmentar la confiança el nombre de regles augmentades disminueix. És interessant comprovar que el nombre de regles de malalties a medicaments generades amb una confiança del 5%, és superior al nombre de regles de medicaments a malalties. Tendència que s'inverteix a partir del 40% de confiança.

Això podria explicar-se en el fet que el nombre d'ítem sets freqüents singleton (aquells ítem sets de només un element) formats per malalties és d'unes 700 mentre que el de medicaments és de 200. També és interessant notar que aquelles regles amb confiança més alta són les que van de medicaments del cor cap a la hipertensió i insulines cap a la diabetis de tipus 2.

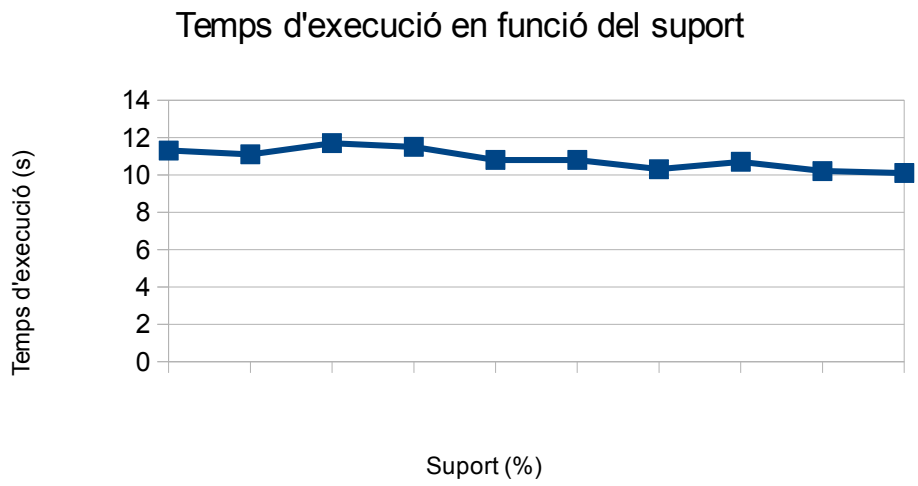


Figura 9. Temps d'execució en funció del suport.

Quan fixem la confiança en un 10% i modifiquem el suport tampoc observem canvis en el temps d'execució, tal com mostra la Figura 9.

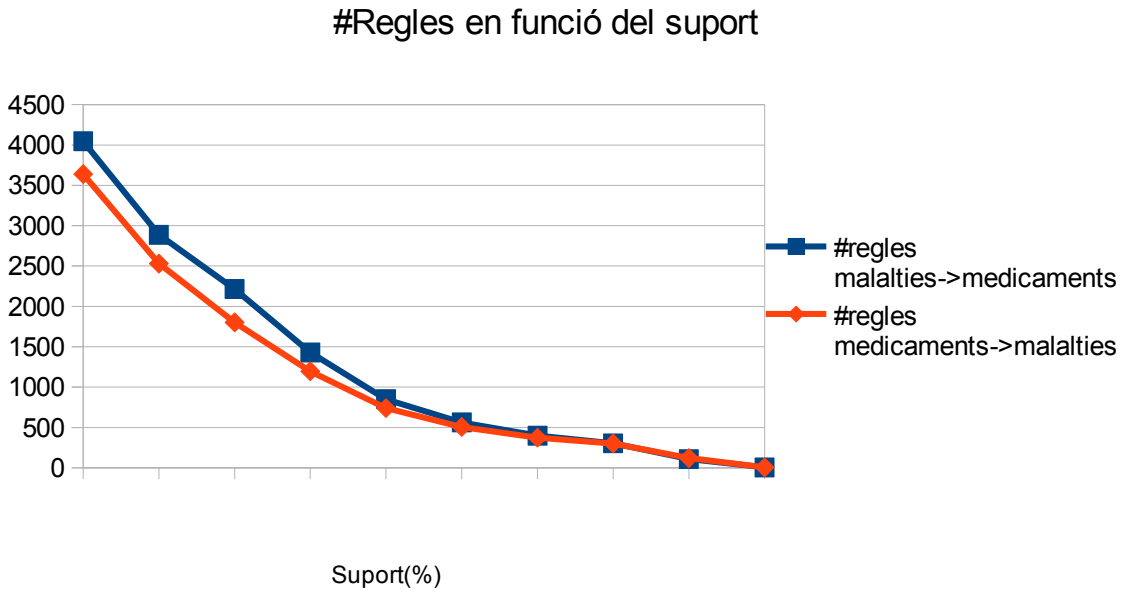


Figura 10. Nombre de regles en funció del suport.

A la Figura 10 observem com a mesura que augmentem el suport el nombre de regles disminueix. La regla amb més suport és la que relaciona la hipertensió amb els trastorns

del metabolisme. Però entre les regles amb més suport és comú trobar-hi les malalties com la hipertensió, la diabetis i la obesitat.

Vistos els resultats es pot esperar que augmentant el suport i la confiança el nombre de regles sigui menor, però les regles que apareixen és poc probable que aportin nou coneixement ja que són àmpliament conegudes. Si es fa disminuir el suport i la confiança apareixeran més regles però serà més difícil de destriar quines són realment interessants.

Avaluació de les regles

Es fa difícil dir si les regles resultants són correctes, ja que caldria que un expert les comprovés o les introduís manualment. Per tant, hem utilitzat com a guia unes taules que ens va proporcionar la doctora Esther Limón que indiquen, per algunes de les malalties més freqüents, quins medicaments s'acostumen a receptar. Aquestes taules es poden tobar en l'Apèndix I i es poden considerar “coneixement expert” que incorporem al nostre sistema. En aquestes taules s'ajunten algunes malalties com els diversos tipus de diabetis (E10, E11, E12, E13, E14), per tant, per fer l'experiment, s'han ajuntat aquests codis en un de sol. En el cas de la diabetis, tots els pacients que tenen E11, E12, E13 o E14 passen a tenir E10. A continuació es mostren algunes de les regles resultants de la diabetis i l'asma tant regles de malalties a medicaments com regles de medicaments a malalties. La resta de regles es poden trobar a l'Apèndix II.

DIABETES MELLITUS (E10, E11, E12, E13, E14)

Regles malalties => medicaments

body	head	Confidence %
E10 DIABETIS	A10BA Biguanidas	56.245
E10 DIABETIS	C10AA ESTATINES	54.798
E10 DIABETIS	A02BC Inhibidores de la bomba de protones	47.592
E10 DIABETIS	B01AC Inhibidores de la agregacion plaquetaria, excluyendo heparina	37.802

Regles medicaments => malalties

body	head	Confidence %
A10AD Combinaciones de insulinas y analogos de accion intermedia y accion rapida para inyeccion	E10 DIABETIS	99.573
A10BG Tiazolidinadionas	E10 DIABETIS	99.314
A10BD Combinaciones de farmacos hipoglucemiantes orales	E10 DIABETIS	99.296
A10AE Insulinas y analogos de accion prolongada para inyeccion	E10 DIABETIS	99.056

ASMA (J 45, J 45.0, J 45.1, J 45.8, J 45.9, J 46)

Regles malalties => medicaments

body	head	Confidence %
J45 ASMA	R03AC Agonistas selectivos de receptores beta-2 adrenergicos	43.146
J45 ASMA	R03AK Adrenergicos y otros agentes contra padecimientos obstructivos de las vias respiratorias	38.008
J45 ASMA	A02BC Inhibidores de la bomba de protones	30.27
J45 ASMA	N02BE Anilidas	21.572

Regles medicaments => malalties

body	head	Confidence %
R03DC Antagonistas del receptor de leucotrienos	J45 ASMA	70.004
R03AK Adrenergicos y otros agentes contra padecimientos obstructivos de las vias respiratorias	J45 ASMA	49.648
R03BA Glucocortocoides	J45 ASMA	48.12
R03AC Agonistas selectivos de receptores beta-2 adrenergicos	J45 ASMA	44.29

S'observa que a les regles de malalties a medicaments apareixen molt sovint les Estatines, l'Omeprazol (Inhibidores de la bomba de protons) i el Paracetamol (Anilidas), ja que s'utilitzen molt sovint i en moltes malalties.

En el cas de la diabetis apareixen en confiança molt alta les insulines en les regles de medicaments a malalties: això és perquè les insulines *només* s'utilitzen en la diabetis.

En ambdós casos els medicaments de les taules que s'utilitzen com a guia apareixen en alguna de les dues llistes. Els altres medicaments que surten acostumen a estar relacionats amb malalties associades. Per exemple, a la taula de la diabetis hi apareixen alguns medicaments relacionats amb la hipertensió; això és perquè molts pacients amb diabetis també tenen hipertensió.

8.2 Grafs

Per fer les proves amb els grafs s'han filtrat les dades de la mateixa manera que en les regles d'associació i a més, s'han descartat les malalties de codis que comencen per Z que són aquelles que realment no són malalties sinó proves que s'han fet els pacients. També s'han descartat aquells pacients que no tenen malalties o medicaments receptats.

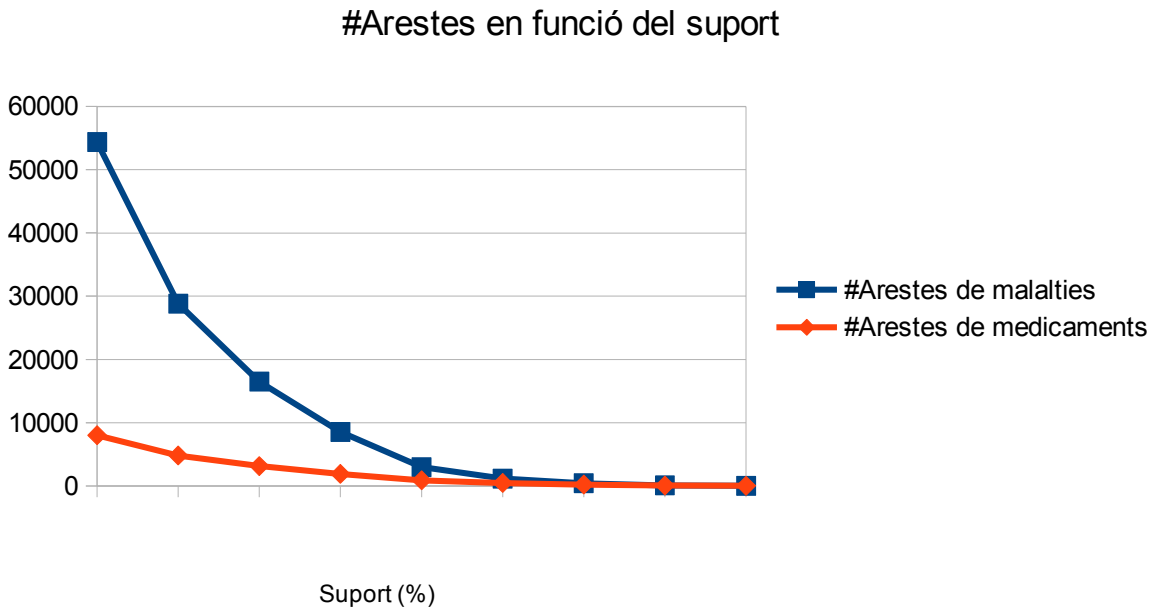


Figura 11. Nombre d'arestes del graf en funció del suport.

A la Figura 11 observem que el nombre d'arestes. Quan el suport és baix, és significativament més alt en el graf de malalties que en el graf de medicaments.

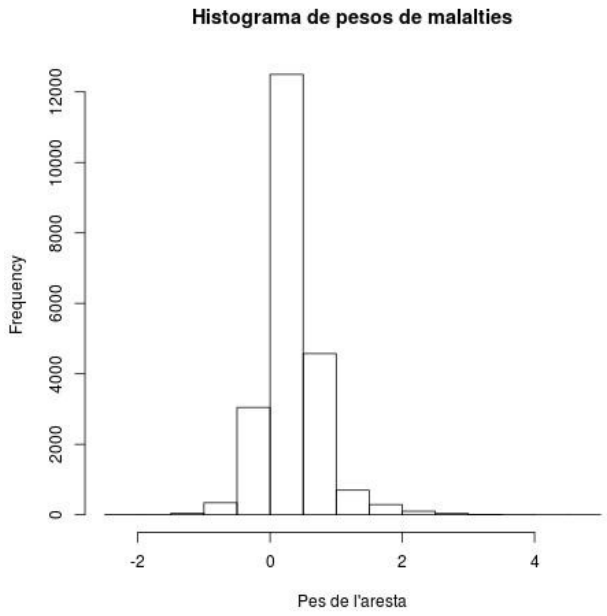


Figura 12. Histograma de pesos del graf de malalties.

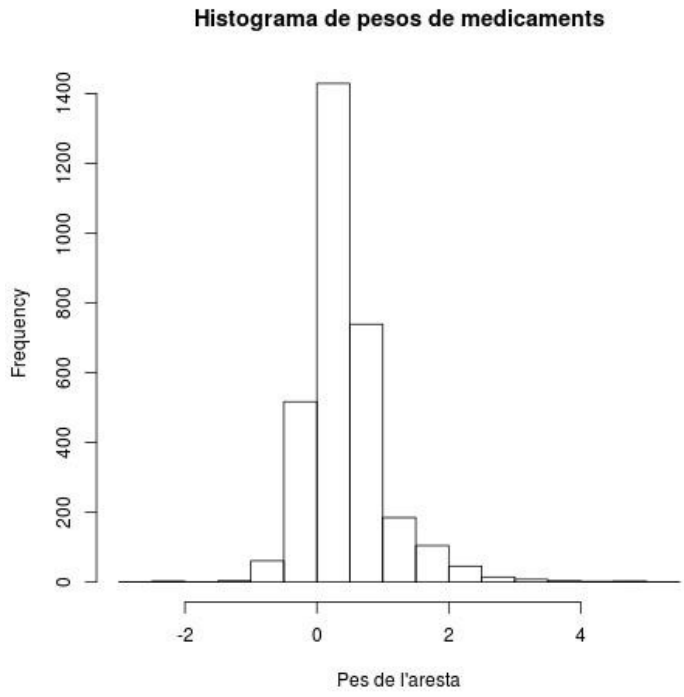


Figura 13. Histograma de pesos del graf de medicaments.

A la Figura 12 i a la Figura 13, es pot veure que la distribució de pesos de les arestes es concentra sobretot quan el pes és proper a 0, és a dir quan la probabilitat de trobar dues malalties o dos medicaments junts és independent. Aquests casos no interessen tot i ser la majoria d'arestes del graf, els casos interessants són els que es troben als extrems.

Això ens dona dues indicacions:

• Quan es facin peticions als grafs s'ha d'evitar que els intervals continguin el nombre 0, per exemple si fem una petició al graf amb l'interval [-1,1] el resultat tindrà milers d'arestes i no podrà ser representat.

• Quan es facin peticions als grafs s'ha d'anar disminuint l'interval més a poc a poc a mesura que ens apropem a 0. Per exemple, si demanem un graf amb un pes mínim de 2 i el graf té poques arestes, podem demanar després, un pes mínim de 1,5 i després si encara en volem més de 1,3 però no disminuir el pes mínim massa ràpidament per evitar que es col·lapsi la interfície.

9 TREBALL FUTUR

A causa de les limitacions de temps d'un treball final de grau, han quedat moltes preguntes per respondre. De fet, com que el projecte tenia caràcter exploratori, es pot dir que encara hi ha més preguntes per resoldre que quan es va començar, però potser a canvi aquestes preguntes també estan més ben definides. En aquesta secció es descriuran algunes propostes que es podrien fer per a estudis i desenvolupaments posteriors. En aquestes propostes els experts de l'ICS hi han contribuït especialment.

Millorar la detecció d'episodis no tancats

Un dels problemes que té l'algorisme per detectar episodis no tancats és que detecta malalties habitualment cròniques com a no medicades. Molts cops però ha sigut una decisió del metge no medicar aquella malaltia segurament perquè la malaltia no es suficientment greu com per a medicar-la. Per exemple, en una hipertensió un metge pot decidir no medicar-la si considera que el pacient no està en risc i posar el pacient a règim per exemple.

La decisió del metge s'acostuma a basar en paràmetres dels quals no s'ha disposat durant l'elaboració del treball, com poden ser el pes o l'índex de massa corporal, el colesterol del pacient, circumstàncies personals del pacient com ara historials previs de reaccions adverses o manca d'efectivitat, etc. Si disposéssim d'aquesta informació es podria crear un classificador fent ús d'algorismes de aprenentatge automàtic. Llavors quan l'algorisme de detecció d'episodis no tancats detectés que una malaltia no es medica, aquests classificadors podrien ajudar a decidir si la decisió és correcta o podria tractar-se d'un episodi no tancat.

Desviacions geogràfiques

Podem dividir les dades per zones geogràfiques i generar regles d'associació entre medicaments i malalties. Podria passar, que en determinades zones geogràfiques les regles per a determinades malalties variessin de forma significativa. En aquests casos s'hauria d'estudiar quines són les causes que els diferents centres mediquin de forma diferent una mateixa malaltia.

Taxonomies

Una taxonomia és una classificació jeràrquica. Podem veure els problemes de salut en forma de taxonomia. Per exemple, la hipertensió essencial pertany a les malalties

hipertensives que al mateix temps pertanyen a les malalties del sistema circulatori. De la mateixa manera podem agrupar els medicaments o fins i tot podem agrupar els centres d'atenció primària per zones.

El problema d'utilitzar taxonomies s'ha enfocat tant per itemsets freqüents [16] com per regles d'associació [17].

Durant l'elaboració d'aquest projecte, els algoritmes de freqüent itemset mining no coneixien les taxonomies dels medicaments o les malalties i manualment s'elegia quin nivell d'agregació s'utilitzava. El problema d'això és que ignorem casos que podrien resultar interessants com podria ser determinats medicaments com l'Omeprazol que concentren gairebé la totalitat del suport del seu grup. En aquests casos ens interessaria més que fos el mateix Omeprazol que aparegués a la regla en comptes del nom del grup agregat.

Utilitzant algoritmes que fossin capaços de discernir aquests casos, es podria millorar la qualitat de les regles donades.

Mesures dels grafs

Un cop s'han generat els grafs de les maneres descrites, és senzill fer diversos càlculs que poden resultar interessants sobre aquests grafs com poden ser mesures de centralitat o influència per conèixer quins nodes tenen un paper més important o aplicar algorismes de clústering o comunitats en grafs per veure quins grups de medicaments i malalties es generen.

Reducció real dels costos sanitaris

Un dels problemes que pot comportar la actual crisi econòmica és que s'intenti reduir la despesa sanitària de manera equivocada. Posem el cas d'un pacient que no es medica correctament com a conseqüència d'una política que intenta reduir els costos farmacèutics de manera cega. Aquest pacient a causa d'una mala medicació ha d'acabar ingressant a un hospital, el cost de tenir-lo ingressat pot superar amb escreix el cost que hagués tingut medicar-lo correctament.

Probablement no es podrà tenir accés als costos hospitalaris, però si que tenim accés als costos farmacèutics. A part de les prescripcions, es disposa de dades que indiquen si un pacient ha passat a recollir a la farmàcia una recepta. També es disposa de les dades relatives a derivacions hospitalàries, és a dir, dades que indiquen quan un pacient a causa de la seva gravetat ha sigut derivat de l'ambulatori a un hospital.

Si es miren les receptes que no han estat recollides a la farmàcia i ho es relacionen amb aquells pacients que han estat derivats a un hospital, es podria saber quins pacients han estat hospitalitzats a causa de no prendre's els medicaments que tenien receptats. Faltaria saber quin és el motiu per el qual el pacient no s'ha pres la medicació i quin cost ha tingut la hospitalització d'aquest pacient. Llavors es podria saber quin és el cost que ha tingut que alguns pacients no s'hagin pogut pagar les receptes.

Suport a l'anàlisi de seguretat sanitària dels pacients

Una de les línies estratègiques del Departament de Salut és la millora de la qualitat i la seguretat del pacient. Uns aspectes importants dins d'aquesta estratègia són la qualitat assistencial i la mesura dels resultats de salut.

En els darrers anys, l'ICS ha desenvolupat diferents eines per facilitar la millora de la qualitat de l'assistència dels seus professionals d'atenció primària (AP). El software d'anàlisi desenvolupat es pot evolucionar cap als aspectes de qualitat i seguretat .

Es disposa de l'Estàndard de qualitat de la prescripció farmacèutica (EQPF) com a indicador sintètic de la qualitat de la prescripció, basat en criteris d'eficàcia i eficiència.

Un aspecte molt relacionat amb l'EQPF és la hiperprescripció, que provoca dues conseqüències: la polimedicació innecessària d'alguns pacients, clarament relacionada amb problemes de seguretat, i la despesa inadequada, que no està relacionada amb la seguretat del pacient però que també és important com a justificació de la intervenció.

Amb el desenvolupament d'aquestes eines, s'espera una millora general progressiva en la majoria d'indicadors clínics utilitzats, tant en la qualitat de l'atenció clínica, com en la qualitat de la prescripció farmacèutica.

Millores en l'escalabilitat del programa

L'arquitectura d'aquesta eina ha estat pensada per tal de poder ser executada remotament en servidors i visualitzada remotament de manera que múltiples investigadors puguin realitzar experiments de forma concurrent i de manera transparent per a ells. No seria molt complicat portar el sistema de fitxers que utilitza el prototip actual a un sistema de fitxers distribuït com pot ser HDFS.

Els algoritmes que s'han utilitzat poden ser paral·lelitzats en gran mesura, de manera que tampoc hauria de ser un problema que el nombre de dades augmentés amb dades d'altres anys i d'altres zones.

Seguretat

Donat que les dades són molt sensibles, caldria millorar la seguretat del sistema per garantir que només els usuaris autoritzats puguin accedir als servidors que les contenen. Ara la seguretat és limitada.

10 CONCLUSIONS

En aquest projecte, s'ha desenvolupat un prototip capaç d'explotar les dades de l'Institut Català de la Salut que pot ser utilitzat per usuaris no experts en informàtica, realitzant experiments de forma remota i que fàcilment pot escalar a més usuaris i més dades.

El software és capaç de detectar quines patologies o medicaments es donen freqüentment alhora en un pacient. Ho fa de forma diferent a com ho feien estudis previs, mitjançant els PMI que a més de ser fàcil d'interpretar pot detectar relacions “negatives”. També relaciona malalties i medicaments de manera que podem saber quins medicaments es donen als pacients amb una determinada malaltia. A partir d'això, el software detecta aquells pacients que se'ls està donant medicaments sense tenir una malaltia que ho expliqui o que tenen malalties sense medicar. Finalment, presenta les dades de forma comprensible i que es pugui treballar amb elles.

Ha faltat però, una manera de validar de forma empírica els resultats que donen els algorismes a causa de que el nombre de resultats és molt gran i no es disposa de cap dataset amb dades etiquetades per tal de poder mesurar la precisió de l'eina. Cal que un expert en medecina es miri manualment els resultats per tal de validar-los. Malgrat això, els resultats que dona són prou coherents. S'han fet servir taules orientatives que ens han proporcionat els experts de l'ICS que ajuden a pensar que els resultats van en la bona direcció.

Com que és el primer cop que es fa un estudi d'aquest tipus i a causa de les limitacions de temps d'un treball final de grau, han quedat molts temes oberts. De fet, a mesura que avançava el treball s'anaven trobant noves àrees que seria interessant investigar, com podria ser intentar quantificar els costos que té per la sanitat intentar estalviar en fàrmacs si després s'acaba produint un ingrés a causa d'un empitjorament de la malaltia del pacient o anàlisis més profunds de les xarxes de medicaments que es generen per el programa.

11 BIBLIOGRAFIA

1: N. Ramakrishnan, D.A. Hanauer, B.J. Keller. “Mining electronic health records.” *Computer*, 2010, 43.10: 77-81.

2: P.B. Jensen, L.J. Jensen, S. Brunak. “Mining electronic health records: towards better research applications and clinical care.” *Nature Reviews Genetics*, 2012, 13.6: 395-405.

3: J. Davis, E. Lantz, D. Page,J. Struyf, P. Peissig, H. Vidaillet, M. Caldwell. “Will this drug give me a heart attack.” In: *the Proceedings of International Conference on Machine Learning (ICML)*. 2008.

4: C. M. Machado, A.T. Freitas, F.M. Couto. “Enrichment analysis applied to disease prognosis.” *J. Biomedical Semantics*, 2013, 4: 21.

5: J. Yang, J. Logan. “A data mining and survey study on diseases associated with paraesophageal hernia.” In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2006. p. 829.

6: I.M. Mullins, M.S. Siadaty, J. Lyman , K. Scully, C.T. Garrett , W.G. Miller, R. Muller , B. Robson, C. Apte , S. Weiss, I. Rigoutsos, D. Platt , S. Cohen, W.A. Knaus. “Data mining and clinical data repositories: Insights from a 667,000 patient data set.” *Computers in biology and medicine*, 2006, 36.12: 1351-1377.

7: D.A. Hanauer, D.R. Rhodes, A.M. Chinnaiyan. “Exploring Clinical Associations Using ‘-Omics’ Based Enrichment Analyses.” PLoS ONE, 2006, 4.4: e5203. doi:10.1371/ journal.pone.0005203

8: H. Cao, M. Markatou, G.B. Melton, M.F. Chiang. “Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics.” In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2005. p. 106.

9: K.I.Goh, M.E. Cusick, D. Valle, B. Childs,M. Vidal,A.L. Barabási. “The human disease network.” *Proceedings of the National Academy of Sciences*, 2007, 104.21: 8685-8690.

10: B. Goethals. “Survey on frequent pattern mining.” *Univ. of Helsinki*, 2003.

11: A. Rajaraman, J. Leskovec, J.D. Ullman. *Mining of massive datasets*. Cambridge University Press, 2011. Capítol 6.

12: T.H. Wonnacott, R.J. Wonnacott. *Introductory statistics*. New York: Wiley, 1972. Capítol 17.

13: G. Bouma. “Normalized (pointwise) mutual information in collocation extraction.” In: *Proceedings of the Biennial GSCL Conference*. 2009. p. 31-40.

14: C.Y. Teng, Y.R. Lin, L.A. Adamic. Lada A. Recipe recommendation using ingredient networks. In: *Proceedings of the 3rd Annual ACM Web Science Conference*. ACM, 2012. p. 298-307.

15: L. Khachiyan, E. Boros, K. Elbassioni, G. Vladimir. “A new algorithm for the hypergraph transversal problem.” In: *Computing and Combinatorics*. Springer Berlin Heidelberg, 2005. p. 767-776.

16: J. Baixeries, G. Casas, J.L. Balcazar. “Frequent sets, sequences, and taxonomies: new, efficient algorithmic proposals.” *Report Number: LSI-00-78-R, El departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Spain*, 2000.

17: R. Srikant, R. Agrawal. “Mining generalized association rules.” In: *VLDB*. 1995. p. 407-419.